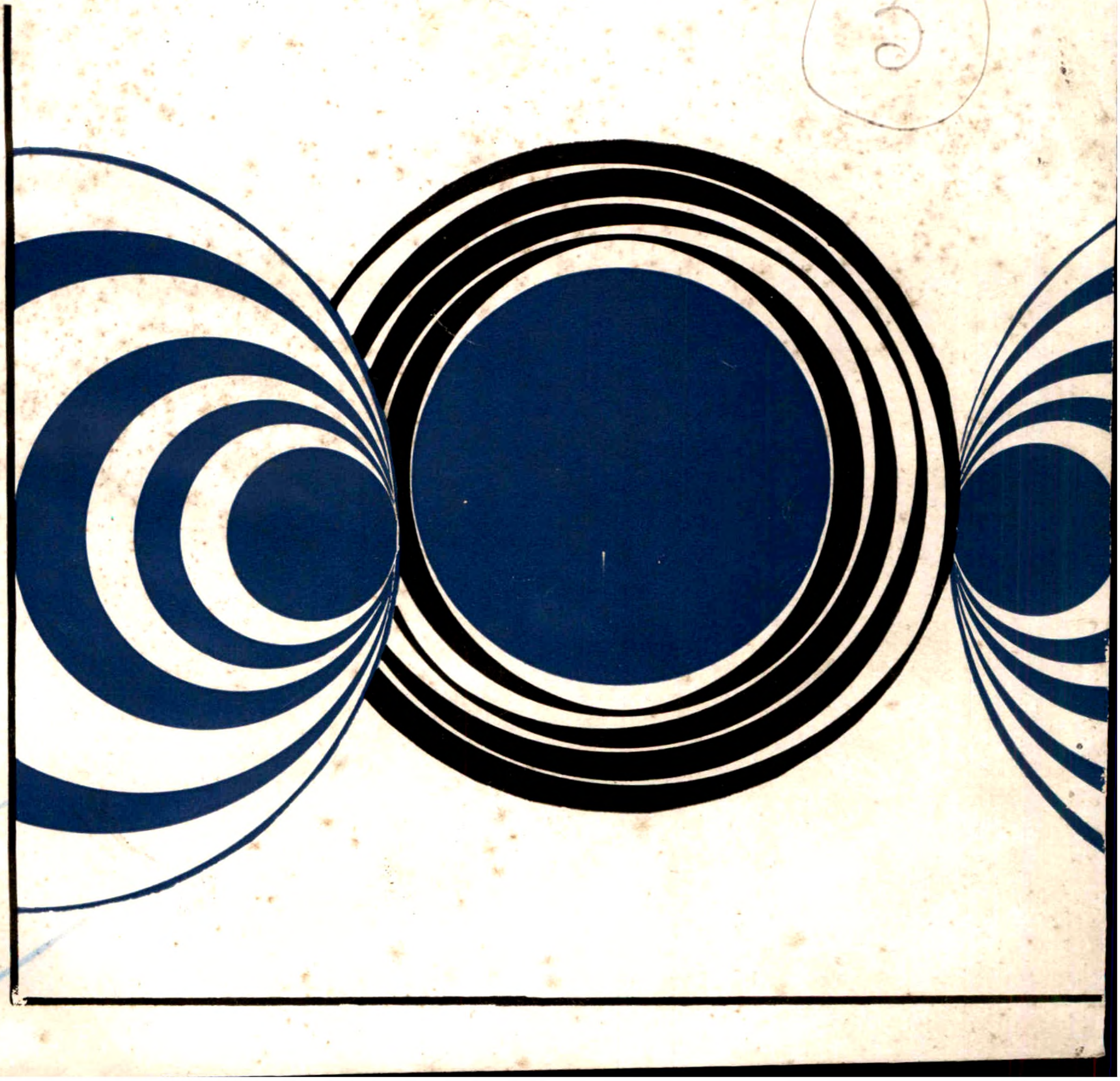
American Documentation

JANUARY
1967
Vol. 18, No. 1

PUBLISHED QUARTERLY BY THE AMERICAN DOCUMENTATION INSTITUTE



AMERICAN DOCUMENTATION

INSTRUCTIONS TO AUTHORS

American Documentation is a publication of the American Documentation Institute. It is a scholarly journal in the various fields in documentation and serves as a forum for discussion and experimentation. Papers already published or in press elsewhere are not acceptable. For each proposed contribution, one original and two copies (in English only) should be mailed to Mr. Arthur W. Elias, Editor, *American Documentation*, Institute for Scientific Information, 325 Chestnut St., Philadelphia, Pennsylvania 19106. The manuscript should be mailed *flat* in a suitable-sized envelope. Graphic materials should be submitted with suitable cardboard backing.

Types of Manuscripts: Three types of contributions are considered for publication: full-length articles, brief communications of 1,000 words or less, and letters to the editor. Letters and brief communications can generally be published sooner than full-length manuscripts. Books, monographs, and reports are accepted for critical review. Two copies should be addressed to the Review Editor, Dr. T. Hines, 54 North Drive, East Brunswick, New Jersey.

Processing: Acknowledgment will be made of receipt of all manuscripts. *American Documentation* employs a reviewing procedure in which all mansucripts are sent to two referees for comment. When both referees have replied, copies of their comments are sent to authors with the Editor's decision as to acceptability. The refereeing procedure requires about 30 days. Authors receive galley proofs with a five-day allowance for corrections. Standard proofreading marks should be employed. Reprint order forms are forwarded with galleys.

Format: All contributions should be typewritten on white bond paper on one side only, leaving about 1.25 inches (or 3 cm) of space around all margins of standard, letter-size (8.5 × 11 inch) paper. Double spacing must be used throughout, including the title page, tables, legends, and references. The first page of the manuscript should carry both the first and last names of all authors, the institutions or organizations with which the authors are affiliated, and notation as to which author should receive the galleys for proofreading. All succeeding pages should carry the last name of the first author in the upper right-hand corner (0.5 inch from the top) and the number of the page.

Style: In general, style should follow the forms given in the Style Manual for Biological Journals (SMBJ), published for the Conference of Biological Editors by the American Institute of Biological Sciences (1964).

Title: The title should be as brief, specific, and descriptive as possible. Vague and unrevealing titles may delay publication.

Abstract: An informative abstract of 200 words or less must be included, typed with double spacing on a separate sheet. This abstract should present the scope of the work, methods, results, and conclusions.

Acknowledgments: Financial support may be listed as a footnote to the title. Credit for materials and technical assistance or advice may be cited in a section headed "Acknowledgments," which should appear at the end of the text. General use of footnotes in the text should be avoided.

Graphic materials: *American Documentation* requires finished artwork. Follow the style in current issues for layout and type faces in tables and figures. A table or figure should be constructed so as to be completely intelligible without further reference to the text. Lengthy tabulations of essentially similar data should be avoided.

Figures should be lettered in black India ink. Charts drawn in India ink should be so executed throughout, with no typewritten material included. Letters and numbers appearing in figures should be distinct and large enough so that no character will be less than 2 mm high after reduction. A line 0.4 mm wide reproduces satisfactorily when reduced by one-half. Graphs, charts, and photographs should be given consecutive figure numbers as they will appear in the text; however, figure numbers and legends should not appear as part of the figure, but should be typed double spaced on a separate sheet of paper. Each figure should be marked *lightly* on the back with the figure number, author's name, complete address, and shortened title of the paper.

For figures, the originals with two clearly legible reproductions (to be sent to referees) should accompany the manuscript. In the case of photographs, three glossy prints are required, preferably 8 × 10 inches.

Organization: In general, papers should state the background and purpose of the study, followed by details of methods, materials, procedures, and equipment. Findings, discussion, and conclusions should appear in that order. Appendixes may be employed where appropriate for extensive lists, statistics, and other supporting data.

Bibliography: Accuracy and adequacy of the references are the responsibility of the author. Therefore, literature cited should be checked carefully with the original publications. References to personal letters, abstracts of verbal reports, and other unedited material may be included. If an as-yet-unpublished paper would be helpful in the evaluation of a manuscript, it is advisable to make a copy of it available to the Editor. When a manuscript is one of a series of papers, the preceding member of the series should be included in literature cited.

Citation Format:

Order: Literature cited should be sequentially numbered as cited.

Authors: Give all authors with arrangement as follows:
Elias, A. W., B. H. Weil, and I. D. Welt

Titles: Give full titles of articles in English, indicating language of original as: (In Ger.)

Journals: Journal titles should be given in full.

Monograph and Serial Data: Should be presented in order as follows: Volume, issue number, pagination, and year. The issue number should be given in parentheses if journal pagination is not continuous from issue to issue. Pagination should be inclusive. Year of publication should be given in parentheses. An example is given below:

Bishop, D., A. L. Milner, and F. W. Roper, Publication Patterns of Scientific Serials, American Documentation, 16 (No. 2): 113–21 (1965).

American Documentation is published in January, April, July, and October. One copy is included in the individual membership fee ($20.00 per year), three copies in the contributing membership fee ($100.00 per year), and up to five copies in the sustaining membership fee ($500.00 per year). Nonmembers may subscribe at $18.50 per year, postpaid in the U.S. Single copies may be purchased for $4.65 each. Communications concerning memberships, subscriptions, reprints, renewals, back issues, advertising, and changes of address should be sent to the American Documentation Institute, 2000 P Street, NW, Washington, D. C. 20036.

American Documentation is indexed in *Library Literature, Current Contents of Space, Electronic & Physical Sciences, Library Science Abstracts, Science Citation Index, Chemical Abstracts,* and *Documentation Abstracts.*

American Documentation is entered for second class mailing at Baltimore, Maryland.

American Documentation

PUBLISHED QUARTERLY BY THE AMERICAN DOCUMENTATION INSTITUTE

Vol. 18, No. 1 JANUARY 1967

Published quarterly by American Documentation Institute and printed by Port City Press, Inc., 1323 Greenwood Road, Baltimore, Maryland 21208. Second class postage paid at Baltimore, Maryland.

Editorial

Each January since becoming Editor, I have used this editorial space to provide a summary of the journal statistics of American Documentation for the prior year. These data are also given in the Annual Report of the Editor to the American Documentation Institute Annual Meeting; but since almost half of the subscribers are not members of the Institute, the information has been repeated. In 1964 and again in 1965 the cover theme of the January issue has been used to provide a graphic display of the publication pattern over the years.

Among the statistics given, a differentiation is made between refereed and non-refereed items. The latter include letters, book reviews, editorials, contents pages, indexes, and the like. Regular journal articles and Brief Communications *are* refereed. The data for 1966 are tabulated below and should be considered in the knowledge that the publication budget was reduced by $4500 transferred to Documentation Abstracts.

Publication Statistics

	1965	1966
Papers	38	29
Pages	381	227
Journal Articles	27	27
Brief Communications	11	2

This brings me to the real subject of the editorial. The unsung referee. Each year I have been increasingly grateful to the anonymous group of gifted people to whom I send manuscripts of a great range of clarity, pertinence and value. I am continually amazed at the care and work devoted by these referees to the brainchildren of others. Calculations, tables, figures, and formulas are checked; bibliographies are updated and corrected; turgid prose is clarified, etc. Even of greater importance a general perspective is provided which enables the Editor to schedule material for publication in a useful and realistic manner.

My files contain dozens of notes and letters from authors expressing gratitude and appreciation for this aid. On occasion acknowledgment is made in the manuscript; often it is not. Remembering that absolutely no compensation is given for these labors, I would like to express my personal appreciation and that of the Institute to the AD referees—professionals in the truest sense of the word. Thank you.

Arthur W. Elias

A Graphic Graphics Card Catalog and Computer Index

The development of a low-cost library control system for visual aids based on punch cards is described. Miniature pictures of the visual aids are printed on the cards with a Xerox 914 Office Copier, identifying information is keypunched, and additional information is entered by typewriter. An inventory of other forms of the visual aid, such as glossy photographs, slides, and VuGraphs, is indicated by check marks on a small form printed on the card. The cards are designed to permit updating by additional punching as well as by overpunching previously punched symbols. The punched data is used by a computer to prepare indexes to the visual aid collection, and the cards themselves are used as a manual card catalog for everyday retrieval of visual aids.

BORIS W. KUVSHINOFF

Applied Physics Laboratory
Johns Hopkins University
Silver Spring, Maryland

The cataloging, storage, and circulation of visual aids is still a rather new responsibility of the Document Library of the Applied Physics Laboratory, and the catalog cards that had been designed initially tended to invite such improvements as inevitably follow from second and third thoughts. For example, ever since we learned of the *Time-Life* picture catalog cards (1), which display a small reproduction of the picture together with descriptive information, we were intrigued with the idea of including pictures of visual aids on our catalog cards.

In the *Time-Life* system the original photograph and descriptive copy (prepared with a typewriter that prints extra large characters) are photographed side-by-side by a Recordak camera, and 3×5 cards are printed photographically from the microfilm. In principle, this system is highly attractive, especially by virtue of its simplicity, but for technical reasons it did not quite fill our needs, as we shall see presently.

The visual aids in question are for the most part 30×40 inch drawings mounted on heavy cardboard (but can be 15×20, 20×30, or even 30×35 inch flip charts). Each of these drawings is usually photographed by the APL Photolab and an 8×10 inch glossy picture, a 2×2 inch slide, a 3¼×4 inch slide, and an 8×10 inch VuGraph may be prepared, depending on the originator's requirements. As added complications, the original visual aid may have one or more overlays, and all or part of these various forms of back-up material may be in black and white, in color, or both. Further, the original and/or overlay may be revised several times. The problem has been, therefore, to find a format in which all of this, as well as other information, could be displayed economically and in a readable manner on a catalog card.

Everything to be known about visual aids, and presumably of use and importance, falls conveniently into two categories, as indicated in the following list:

1

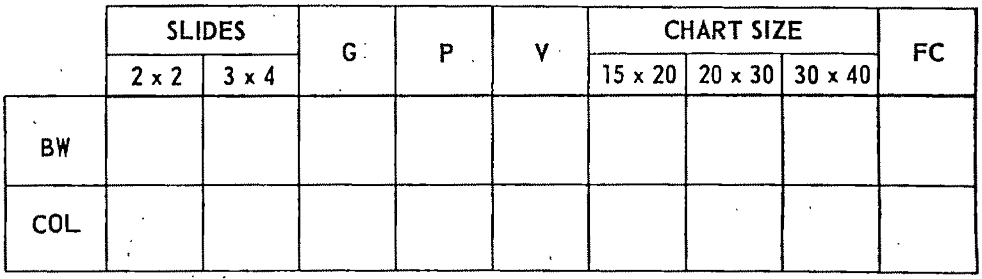
Visual Aid and Inventory
of Back-Up Material
Physical Description of

Chart Size
Mounted Chart
Flip Chart
2×2 Slide
3×4 Slide
VuGraph
Overlay
Print
Glossy Photo
} Black and White or Color

2

Historical Information and
Identification of Visual Aid

Title of Chart
Originating Group or Project
Date of Chart
Originator's Number
Originator's Name
Artist's Name
Presentation for Which Chart Was Prepared:
  Date of Presentation
  To Whom Given
  Where Given
  Security Classification
  Charts Used
Photolab Negative Number
Security Classification of Chart
Revisions



	SLIDES		G	P	V	CHART SIZE			FC
	2 x 2	3 x 4				15 x 20	20 x 30	30 x 40	
BW									
COL									

FIG. 1. Form designed to show inventory of back-up material and physical description of original visual aid. "G" stands for glossy photograph; "P" for Ozalid, Multilith, or Xerox print as applicable; "V" stands for VuGraph; and "FC" for flip chart, which is always 30 × 35 and therefore needs no size breakdown.

The solution to the problem posed by the first column almost suggested itself, and proved easy to enter, compact, and readable. The small form reproduced here is filled out as necessary and is marvelously simple and effective.

This form takes little space on the catalog card, and yet, by putting X's in the applicable squares, takes care of all the possible combinations that can exist for a visual aid and its back-up material.

The problem posed by the information in column 2 of the list was not so easy to cope with. Titles can be long, there may be subtitles, and considerable space can be taken up by each of the other items in the list. This clearly indicated the need for some shorthand or coding system. At this point in the development of the improved visual aid (VA) system, two factors came into play, one a matter of personal preference, the other an out-and-out misconception. Were it not for this misguided good fortune, the trail may not have led where it did.

The original visual aid catalog consisted of 5 × 8 cards, which provided for most of the information given in the list and allowed for several subjects. To one unfamiliar with the format on these cards, the arrangement of the information seemed to be not as effective as it might be. Being predisposed to a catalog card format in which the information is printed high on the card (to facilitate easier reading in well-filled drawers), the rearrangement tended to reposition the required information upward, so that we got a configuration as shown in Fig. 2. The misconception was that we thought the negatives made by the photolab were $3\frac{1}{4} \times 4\frac{1}{4}$ (instead of 4 × 5 and 8 × 10, as they actually are). Mindful of costs, we wanted to avoid the expense of rephotographing the entire collection. As it turned out, $3\frac{1}{4} \times 4$ slides are available for a large part of the visual aid collection, quite a convenience for the diazo process, which works most simply with a positive transparency.[1]

By reducing some of the desired information into coded form, repositioning it high on the card, allowing space to the right for a 3 × 4 picture, we were left with considerable blank space at the bottom.

Then, almost by intuition, we realized that if we got rid of all the blank space at the bottom of the card, we had a shape very similar to that of a punch card (outlined area in the figure). This idea, quite naturally and directly, led to all sorts of other ideas. If the catalog cards could be made of stiff photographic paper, and if the cards could be keypunched and fed to a computer for indexing purposes, we would indeed have a versatile system. Not only could we have a manual card catalog, we could provide our users with a variety of computer-printed indexes of the entire collection.

There were two methods immediately available to us for putting pictures on cards. One was Xerography, the other photography. Since the APL Photolab already had negatives of one size or another, we decided to explore the photographic technique first.

Knowing that typing on photographic emulsion is impractical, we wrote to Eastman Kodak[2] to see if we could die-cut, dimensionally stable cards about 6.5 mils thick, with photographic emulsion on one half and bare paper on the other. The answer was that this was technically feasible, but exorbitant in cost for a small market. Dimensional stability is enormously difficult to maintain because photographic paper is wetted during development and fixing, and usually is dried by heat. Furthermore, emulsion and paper have different hygroscopic characteristics and different coefficients of expansion. This, compounded by core curl that often sets during manufacture, pretty much eliminates ordinary photographic paper, as we know it, for use as punch cards. One positive benefit of the exchange with Eastman Kodak was their suggestion of a typewriter ribbon that would "take" on emulsion and the suggestion that a clear acrylic spray could be used to eliminate rub-off.

Meanwhile, approaching the problem from another angle, we evolved a photosensitive diazo punch card (Fig. 3). To test the idea, we stimulated the inventiveness of two friends[3] in the Laboratory, and with their

[1] Somewhat surprisingly, black and white slides of line drawings reproduce quite well on the Xerox 914 Office Copier.

[2] Thanks are extended to Mr. John P. Eager, Vice President of Recordak, and Mr. H. H. Haggedorn, Product Specialist, for several useful suggestions.

[3] L. W. Fraser and Henry Jorgensen, Fleet Systems Division, APL.

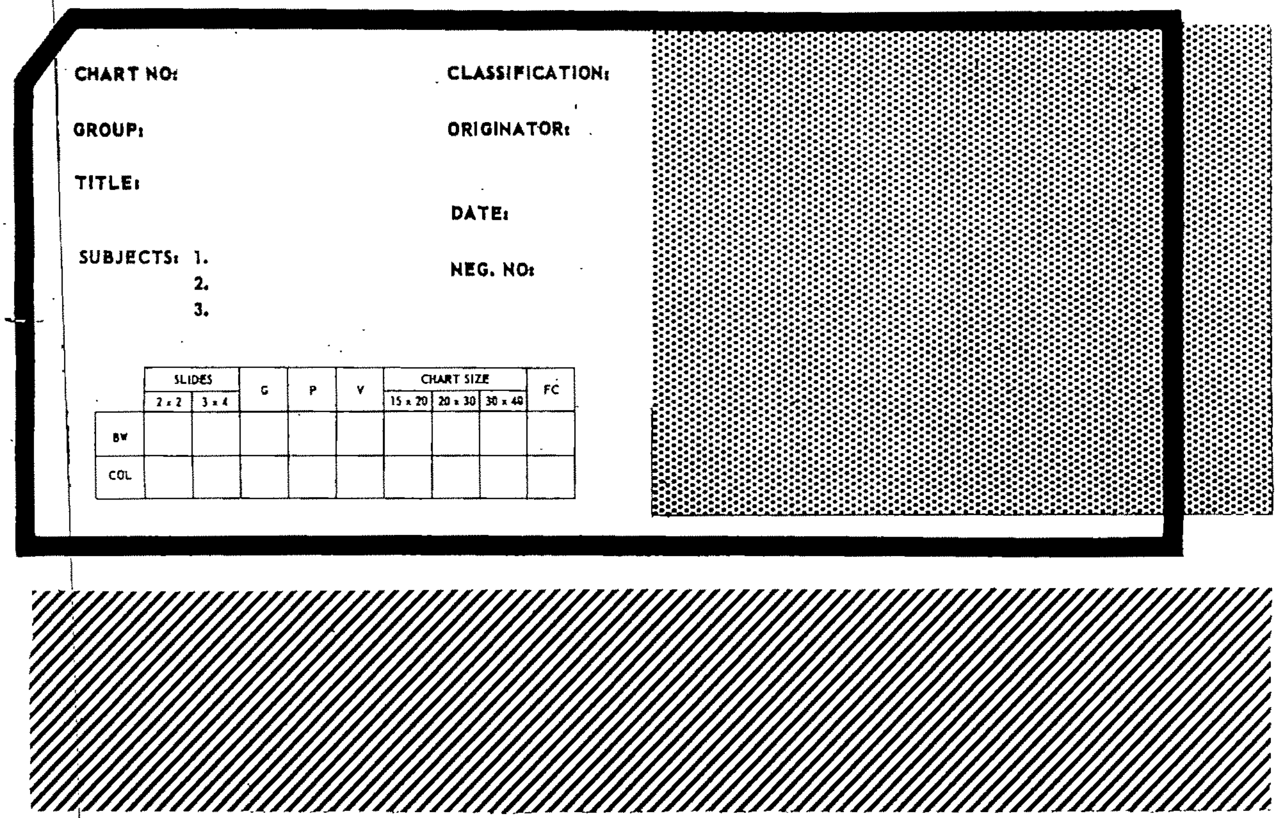


CHART NO:

CLASSIFICATION:

GROUP:

ORIGINATOR:

TITLE:

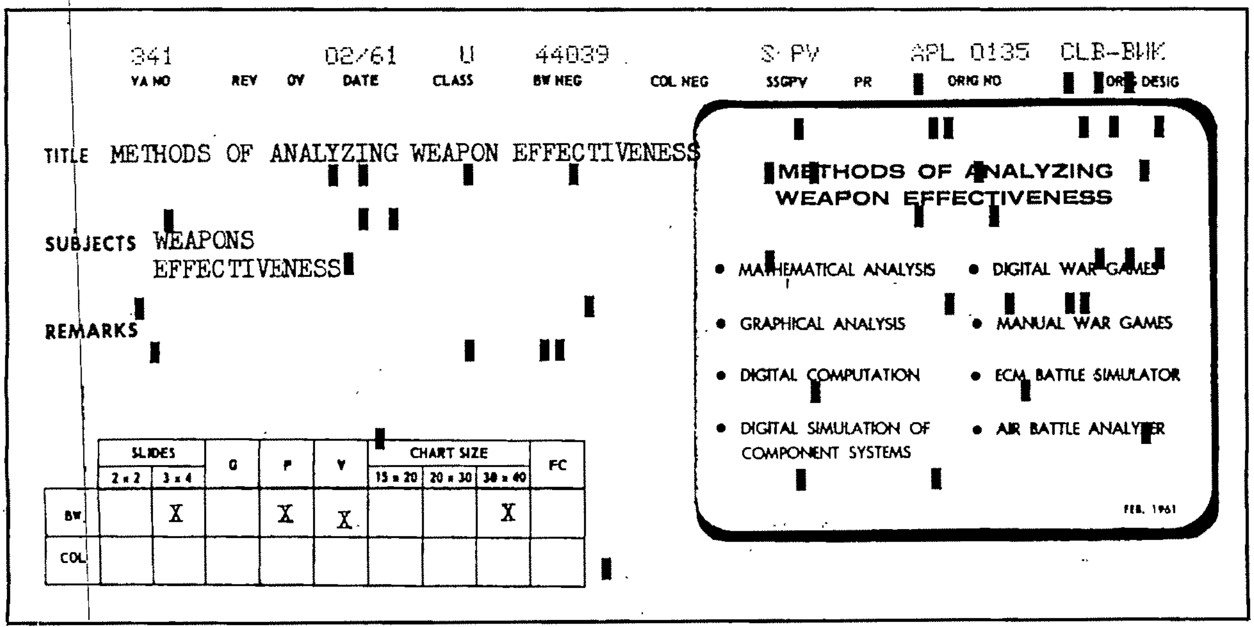
DATE:

SUBJECTS: 1.
2.
3.

NEG. NO:

	SLIDES		G	P	V	CHART SIZE			FC
	2 x 2	3 x 4				15 x 20	20 x 30	30 x 40	
BW									
COL									

FIG. 2. Positioning of information on catalog card. Shaded area is reserved for picture, hatched area at the bottom is waste space. Black outline shows relative size of a standard tab card.



341			02/61	U	44039		S PV		APL 0135	CLB-BHK
VA NO	REV	OV	DATE	CLASS	BW NEG	COL NEG	SSGPV	PR	ORIG NO	ORIG DESIG

TITLE METHODS OF ANALYZING WEAPON EFFECTIVENESS

SUBJECTS WEAPONS
EFFECTIVENESS

REMARKS

	SLIDES		G	P	V	CHART SIZE			FC
	2 x 2	3 x 4				15 x 20	20 x 30	30 x 40	
BW		X		X	X			X	
COL									

METHODS OF ANALYZING
WEAPON EFFECTIVENESS

- MATHEMATICAL ANALYSIS
- GRAPHICAL ANALYSIS
- DIGITAL COMPUTATION
- DIGITAL SIMULATION OF COMPONENT SYSTEMS
- DIGITAL WAR GAMES
- MANUAL WAR GAMES
- ECM BATTLE SIMULATOR
- AIR BATTLE ANALYZER

FEB. 1961

FIG. 3. Final version of the Graphic Graphics Catalog Card

assistance hand-manufactured a few cards to determine if they would feed through a keypunch machine and subsequently act as inputs to the IBM 7094 Computer. Using 102-Z Ozalid paper, a product that had recently become available and which is plastic coated on the back, we succeeded in making a productive run, even though the cards were inadvertently trimmed slightly smaller than standard punch cards.

Briefly, the cards were made as follows: We copied a 3¼×4 glass slide on diazo film, made a positive transparency of a paste-up of the standard matter that we wanted on each card, and stripped the two pieces of film together. We were now in business to print our sample cards.

The test cards appeared much like the final version shown in Fig. 3, and despite the fact that special care was not taken to achieve quality, everything on the test cards was readable. The title and subjects were typed on a Synchrotape,[4] and the typed line across the top of the card was produced by the IBM 26 keypunch machine. The few holes that appear in the picture are not at all troublesome, since the main purpose of the illustration is for recognition of the original.

Before discussing one or two of the more interesting features of the system, let us briefly review its automatic indexing capabilities. Most of the fields on the card, shown in Fig. 3, are self-explanatory, but some deserve additional comment.

From the outset, we tried to maintain a capability of updating the cards because it would be costly in both time and money to have to produce a new set of cards just to change a security classification, for example, or to indicate a revision. Therefore we used a field of three columns for the security classification, the first of these for "secret," the second for "confidential," and the third for "unclassified." If a secret chart is downgraded, a C is entered in the column next to the S, and if the chart is completely declassified, a U is entered next to the C. Thus, the symbol farthest to the right on a card is the correct classification of the item. The column directly to the left of this field is used to indicate the security classification downgrading group number, and the column directly to the right is used to enter a symbol indicating proprietary information, if applicable. Five columns are allowed for the various types of back-up material. Thus, if a glossy print or a slide is made up some time after the cards have been filed in the catalog, the cards can be pulled and additional punches made in the card. The only difficulty in reading this field may be in distinguishing the first S (standing for 2×2 slide) from the second S (standing for 3¼×4 slide). But one need only glance down at the inventory form in the lower left corner to ascertain which S it is; furthermore, in the indexes there will be neighboring entries that will dispel any possible confusion.

The Revision and Overlay fields are used in conjunction with the VA number field. A separate card is made up for each overlay, which is given the same number as its associated chart, but with an alphabetical suffix to indicate that it is an overlay. An original overlay is indicated by −A, its first revision by −B, etc. Revisions of charts are indicated by numerical suffixes.

Assume that a chart numbered CH 543 is the 123rd chart to enter the VA system; the chart is assigned the number 123. Now assume that the chart is revised. The catalog card numbered 123 is updated by entering —1 in the REV field, and an asterisk is punched after the number 123, i.e., 123*. After being added to the computer index tape, the card is returned to the catalog to represent all of the back-up material that has not been revised. Meanwhile, the revised chart numbered CH 543 (now VA 123–1) is photographed by the photolab, and a new card is made up by the Document Library. This new card is filled out and keypunched as usual and reproduced. The keypunched card is added to the master index tape and then filed. We now have two sets of cards related to one basic chart, one showing the way it *was*—the other showing the way it *is*.

Now, if the same chart is revised again, the new version is numbered 123–2, and a new card is made up and processed as before. Meanwhile, the old cards are updated by overpunching the hyphen in the REV field of the first card with an asterisk and doing the same in the VA field of the second card, and adding —2 in the REV field of the second card. These steps are shown in sequence in Fig. 4.

Thus, the hyphen in the VA field indicates that the version of the original visual aid as displayed on the card still exists, and an asterisk indicates that it has been revised or intentionally destroyed. The hyphen and asterisk are used because cards do not have to be remade; an asterisk can be punched over a hyphen, thereby updating the card and indexes by a single stroke.

The OV (overlay) field is used in exactly the same way as the REV field, except that letters —B, —C, —D, etc., are used instead of digits. Complete sets of cards are made up for overlays just as for the charts themselves, the original overlay being numbered, for example, 123-A, simply to indicate that it is an overlay and to distinguish it from a chart.

The last two fields on the card are reserved for the originator. He is free to use them in any way he wishes to arrange his particular set of visuals in a systematic manner. As long as he uses a system, his set of visual aids will be grouped together in the index. A sample of the VA index is shown in Fig. 5.

As currently planned, there will be four indexes: (1) by VA number; (2) by negative number; (3) by originator's designation; and (4) by originator's number. After a master card is made up, filled in, and keypunched, five copies will be reproduced by Xerox on Xerograkards: three or four will be filed by subject, one will be presented to the originator for his use, and the master will be filed

[4] A Synchrotape was used because it had small type, and we needed several cards for testing; actually, any typewriter may be used.

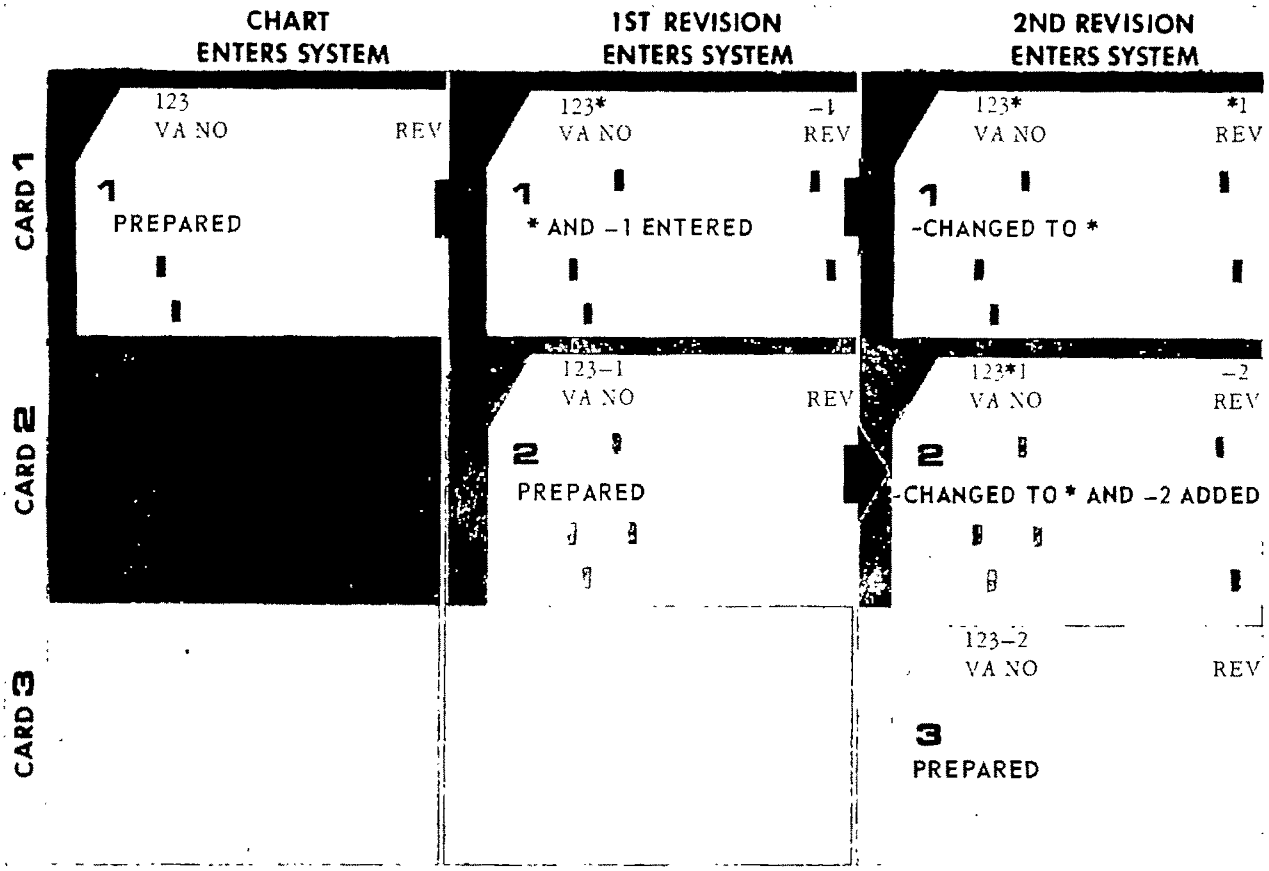
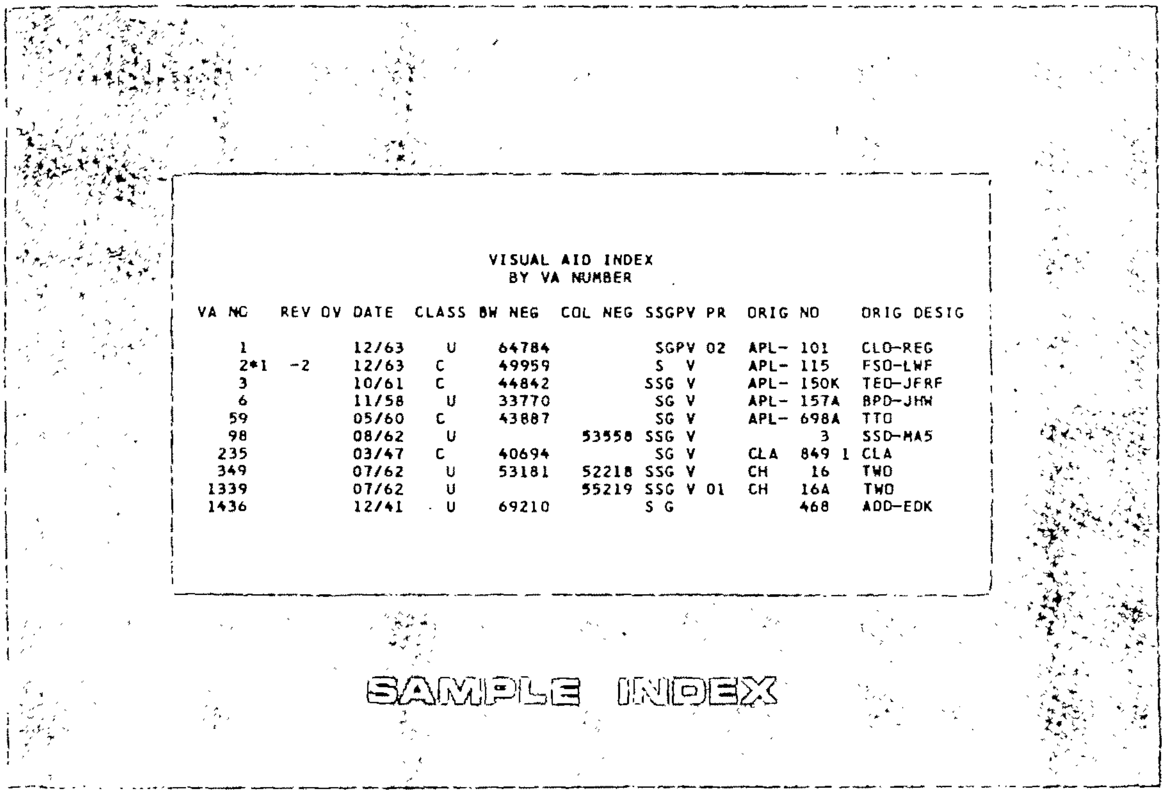


Fig. 4. Sequence of steps for updating revisions



```
                         VISUAL AID INDEX
                          BY VA NUMBER

 VA NC   REV OV DATE  CLASS BW NEG  COL NEG SSGPV PR  ORIG NO    ORIG DESIG

     1          12/63    U   64784            SGPV 02 APL- 101   CLO-REG
     2*1  -2    12/63   C    49959            S   V   APL- 115   FSO-LWF
     3          10/61   C    44842           SSG V    APL- 150K  TED-JFRF
     6          11/58    U   33770            SG V    APL- 157A  BPD-JHW
    59          05/60   C    43887            SG V    APL- 698A  TTO
    98          08/62    U            53558 SSG V             3  SSD-MA5
   235          03/47   C    40694            SG V    CLA  849 1 CLA
   349          07/62    U   53181    52218 SSG V     CH    16   TWO
  1339          07/62    U            55219 SSG V 01  CH    16A  TWO
  1436          12/41  . U   69210          S G             468  ADD-EDK
```

Fig. 5. Sample of visual aids index

by VA number after it has been passed through the computer for the indexes.

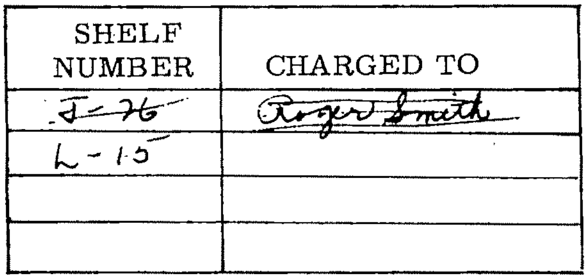
One feature of these cards is especially appealing from the user's standpoint. He need not use the cumbersome originals or even slides, VuGraphs, or glossy prints to organize a presentation. Transparencies must be viewed by transmitted light, and 8×10 glossy prints are too large to spread out in any number for comfortable simultaneous examination. The cards, on the other hand, are not only compact, but carry all of the information to be known about the visual aid, in addition to a small picture of the original itself. A staff member preparing a presentation can spread out as many as 30 to 40 of these cards on his desk, make his selection, choose the sequence of presentation, and then send the cards, in the order desired, to the Document Library. The Document Library can than arrange slides or VuGraphs in the same order, place them in a carrying box, and charge them out to the requester.

At the presentation, the staff member could conceivably use the cards as cue cards and might even be able to get along without a typed speech. In any event, he will know exactly what is coming up next, what it looks like, what the title is, etc., by flipping the cards one by one as he proceeds through his talk. He need not crane his neck to see the screen while he talks.

The storage and circulation system is the essence of simplicity. The backs of the cards are printed as shown in Fig. 6. The charts themselves are stored in several slotted cabinets labeled A, B, C, etc. The slots in each cabinet are numbered 1, 2, ... , 30. The shelf location of each chart is indicated by the appropriate combination of cabinet letter and slot number.

The cabinet slots are provided with rollers, and the edges of the charts (stored back to back, two to a slot) are protected with tight-fitting metal channels running along the 40-inch edges. Since the charts are stored back to back, it is easy to determine which of the two is desired, simply by looking at them. If a chart is circulated, the borrower's name is written on the back of the card, thereby becoming a permanent record.

The last problem involving circulation was solved so easily that we have forgotten why we thought it serious. If someone requests charts, slides, and VuGraphs at one time and returns them at different times, how do we check these items in and out quickly without a lot of writing and record-keeping? Our solution, certainly not the only one, is to slip the appropriate card of the VA file into a transparent sleeve. A marking pencil is used to circle the appropriate boxes of the inventory form. When some items are returned but others not, the circle around the returned item is wiped off or scribbled out. When everything is returned, the card is removed from the sleeve and the sleeve is discarded (if it is disposable), or wiped clean and used again (if it is of sturdy acetate). These sleeves cause the card to stand up a fraction of an inch above the other cards in the file, which is a handy feature in itself.



SHELF NUMBER	CHARGED TO
~~J-26~~	~~Roger Smith~~
L-15	

FIG. 6. Charge-out and storage location record

We have already proved that the diazo process can be used. Two factors make it especially attractive: the 102-Z Ozalid paper appears to have the dimensional stability and flatness to serve for punch cards even after rather abusive handling. Material as well as equipment costs are about as moderate as one could possibly hope for, and quality of the end-product is very good, especially if halftones are involved. Although we are testing other photosensitive papers, the plasticized diazo paper commercially available seems best able to satisfy all minimum requirements of our system.

The greatest drawback of the diazo process in this application is its slowness. Xerograkards,[5] which are perforated punch cards, four to a sheet, designed for use with the 914 Xerox Office Copier, require much less time to produce.

With these cards we found we need to type, punch, and print only one card, and reproduce six copies of four different cards at a time in a little over half a minute, as compared to something like 20 minutes by diazo.

The necessary copies of the cards can be produced in various ways, but the first step in each case is to make a reproducible master card that has all the information on it, including the picture. Two methods appear practical in our case:

1. We offset-print a quantity of perforated sheets with standing material (including the charge-out form on the back). To produce the master cards, we take photographic prints of the right size, position them in a jig, and Xerox one copy on the preprinted sheets. Since there are four cards to a sheet, we can make up cards for four charts at once. These cards are torn apart, typed, keypunched, marked as necessary, and re-Xeroxed five times on plain sheets for the remaining cards needed. Since this means Xeroxing a Xerox copy, we have to be content with background and haze on the cards. Also, very small type is just barely readable by magnifying glass.
2. The second method is to produce the master card on diazo paper. The quality of a diazo master card is considerably higher in comparison with Xerox, and better quality Xerox copies result. Furthermore, halftone images and images with large solid

[5] Trademark of Busiforms, P. O. Box 34086, San Francisco, Calif. 94134.

areas come up much better on diazo. It must be noted, however, that the diazo process is tedious and time-consuming as compared to the Xerox method.

Anyone who has reached this point and remembers that we have not provided for the illustrator's name on the card, must be satisfied with this explanation: we have passed this responsibility to the originator, who has two 10-column fields in which to put this information if he so desires. The illustrator's name is sometimes quite important, for having done the original work, he requires much less instruction and information to make quick changes in updating the visual aid.

Other applications of the graphic punch card have been explored in a superficial way. The system could certainly find use in picture libraries, in art museums, and possibly in newspaper libraries. These cards are readily machine sortable. There may be very real applications for them in medical research (2) and perhaps even in criminology in connection with fingerprint identification and the "mug" file. It would be extremely interesting to see if graphic tab cards could be used in identifying and classifying satellite weather photographs. Anyone having access to a computer with video and optical capabilities might find it interesting to explore the possibility of coating the back or edge of a graphic card with magnetic oxide. Graphic images and other forms of data could be stored magnetically on such cards. These cards could be processed by magnetic reading heads as well as by optical scanners. One likely application would seem to be for parts catalogs (3). Descriptive material and pictures of the parts shown on the card could be used visually by editors. The cards could then be fed to an automatic composing machine.

Estimated cost for programming and for IBM 7094 computer time for the preparation of four indexes for a collection of 5000 visual aids is $900. Printing 30 copies of a 300-page index is estimated at $60, and the cost of the perforated Xerograkards is $376 (10 M sheets, 40,000 cards).

The over-all cost is thus less than $1,400. Adding approximately $500 for Xeroxing and miscellaneous items, we should be operational with this collection for about $2,000, or 40 cents per visual aid. Salaries and overhead are not included. These rough figures are provided merely to give a general idea of what such a system might cost.

References

1. *Photo Methods for Industry:* 55–56 (October 1960).
2. Auerback, S., and A. Lovitz, Jr., Electronic Representation of Histological Patterns, Part I, *Journal of Laboratory and Clinical Medicine* (December 1965).
3. Maurer, W. H., and M. L. Reynolds, A Program System for the Automatic Issue and Revision of Illustrated Parts Catalogs, *IEEE,* Vol. EICI–11 (No. 1):71–78.

The Use of Second Order Descriptors for Document Retrieval*

It is proposed that one way to increase the efficacy of document retrieval is to define to the computer the descriptors used to index the file. A computer program written in COMIT to implement the proposal and to facilitate testing its capabilities is described. Definitions are given to the computer as a string of terms called a "definitor." These terms, which act as "second order descriptors," are not normally those used as file descriptors. Their introduction provides a controllably broader base for link-finding and matchcounting operations by the computer. It also makes possible such things as introducing new terminology and biasing existing descriptor indexes towards special interests or languages without having to re-index the file. The program computes a "pseudometric distance" between a query and each document and prints an ordered list of those documents closer to the query than some chosen cut-off value. (Large files would probably require some preselection, such as that which would result from use of a concordance.) It then substitutes for each descriptor its definitor and repeats the above process. The result is that the subjective human judgment required to evaluate the efficacy of introducing the definitors is reduced to a statement as to which list would be considered more useful. Use of the program to date has been only as a demonstration so no conclusions can be stated other than that the demonstration results would seem to indicate that testing on a serious scale should be undertaken. (This paper is a result of work sponsored by The MITRE Corporation's Educational Assistance Program.)

MILES A. LIBBEY

Director
Information Planning Program
American Institute of Physics

1. Introduction

It is general practice today to use strings of natural language terms to represent to computers the conceptual contents of documents. This is called "coordinate indexing" and the terms used, whether single words or simple phrases, are usually called "descriptors." Although the debate continues as to the advantages and disadvantages of coordinate indexing compared to other techniques, the fact is that essentially the only access today to the conceptual content of a truly large number of, if not most, documents of current scientific and technical interest is by means of their assigned descriptors. If access by automatic means is stipulated, then these descriptors constitute almost the sole entry to that literature in machine-readable form. Not only is a large part of the literature *already* entrusted to descriptors for safekeeping, the rate of addition to this store is increasing both numerically and proportionally. Considering the vast investment of resources (including but not limited to money) in the research and development results so stored and the vast investment in current R & D in progress which need to be related to that store, it is an urgent matter to do anything within reason that would offer some hope of wringing greater utility from those descriptors which have already been assigned. This is in addition, of course, to continuing our efforts to learn how to do better indexing in the first place. The proposal I am making concerns itself directly only with getting automated information retrieval systems to make better use of descriptors that have already been assigned. However, I feel sure that it would have implications for the descriptor assignment process in any information system which was using it, and would interact therewith in a beneficial and complementary fashion.

* A condensed presentation of this paper was given to the 31st Congress of the International Federation for Documentation in Washington, D. C., on October 14, 1965.

• 2. Fundamental Considerations

The rationale for my proposal requires some consideration of fundamental principles of coordinate indexing. I assume that it is generally accepted that some degree of terminology control is required for good coordinate indexing. B. C. Vickery has neatly summarized the general view as to how this is done as follows: "The control of terms for use as descriptors is essentially a matter of establishing relations between words." (1)

This statement reflects the viewpoint, apparently now universally held, that an important — even characteristic — part of the process of indexing a document is deciding for all time just what aspects of its substantive contents may some day be of interest. (I use the word "document" in its broadest sense.) Now, obviously, if a document is about guns and butter, and the indexer only mentions guns, the fact that it was also about butter will be lost to posterity as far as access through that index is concerned. But it is usually assumed that such decisions will be made correctly. What the bulk of the literature on the specifics of coordinate indexing (and indexing in general) is about is how to decide, during the indexing process, what relations to show and how to show them. In general this is reflected in Mr. Vickery's statement and the term "establishing" furthers a connotation of finality in the fixing of some decision or the consummation of some act.

But what would happen if, in the indexing process, a document on poodles did not get indexed also by the term "dog?" Would this mean that the chance to use the generic relation of "dog" to "poodle" has been lost forever? There seems to be no reason, at least in principle, why it should. Regardless of the acts of any indexer, a poodle is still a kind of a dog and will stay such. There are other terms that "poodle" implies — or *could* according to one's interests imply: "animal," "pet," "a canned dog food," "clipping parlor," etc. Presumably, a human acting as an intermediary between a "user" and an information store knows these things and makes use of them, either consciously or subconsciously, in deciding how to approach the information store to best help the user. That is, the chasm between the terms in which the user expresses his request or query and the terms in which the contents of the information store is expressed is bridged by the capability of the human intermediary to relate these sets of terms through his knowledge of their meaning. The principal component of this knowledge is surely derived from the characteristics of a natural language, both general and specific, and in the ability of a native speaker to exploit them, both consciously and subconsciously. Other components undoubtedly derive from knowledge of the information store itself, of the subject matter, and perhaps of the personal needs or viewpoints of the user.

But the time is about here when direct access by users (which may in some cases be other automated systems!) is the order of the day. Is there a way to include in the automated system any effective substitute for the knowledge of meaning previously supplied by the human intermediary? I think there is.

The relations between words that we need to use in the process of retrieving desired information from the store are *not* relations that some indexer might — or might not — have "established." They are, rather, relations which exist intrinsically, inherently, and potentially in terms by virtue of their being part of a natural language! The important implication of this is that *to the extent* that descriptors are used in their natural language sense, those relations, being inherent in them, remain just as available for selection and exploitation during the retrieval process as they are for the original indexing process. For example, "poodle" still implies "dog." Furthermore, since there is no reason for limiting ourselves to one or two relations of special concern, we can, in principle at least, list relations in sufficient number and detail, and, if we choose, in such a holistic fashion, as to effectively approach the substantive or referential effect of a definition, i.e., of an indication of meaning.

The problem then becomes one of practicability. How can this "inherency" of relations be utilized to enable the selection and exploitation during the retrieval process of relations of special interest or to provide some effective substitute for that knowledge of meaning heretofore contributed by a human intermediary, or, better yet, both? My proposal for doing this combines the complementary powers of what I call a "definitor" with those of a quantifying and normalizing function such as a "pseudometric." These are explained in the next two sections.

• 3. The Definitor

By a "definitor" I mean a string of terms that serves the computer as a surrogate for a definition, defining to it the descriptors used to index the file. The definitor provides the mechanism for making any of those relations that are inherent in terms explicit and accessible, hence manipulable. It simultaneously provides a vehicle for carrying in some useful sense information about the meaning of those terms entering into the retrieval operation, previously provided by the human intermediary.

Each constituent term of the definitor will, as I currently envision it, be a single term, symbol, or simple phrase. Taken together, as a definitor, these terms are used to characterize the meaning of a file descriptor in much the same manner as the string of file descriptors is used to characterize the substantive contents of the document they were assigned to index. They could be used to retrieve, relate, and compare terms just as the file descriptors are used to retrieve, relate, and compare documents, and in fact I have so used them. (2) For the present purposes, however, definitors are used to replace the file descriptors in the "description" of a document. Such replacement results in the definitor constituents assuming a new role of acting directly as new document descriptors. I call these "second order descriptors" to refer to this role and yet retain the distinction between these and the original file descriptors. Correspondingly, I sometimes refer to the file descriptors as "first order descriptors" when comparisons are being discussed.

These second order descriptors usually would not be those used as file descriptors. As far as I can see, however, it will always be advantageous to include a repetition of the file descriptor as one of the constituents of its own definitor. As will be seen later, we will be looking for matches, so why throw away possibilities for direct matches?

Second order descriptors would not necessarily be taken from any kind of a terminology control list. I would expect that in practice most of them will come from various sources, most of which could be construed as terminology control lists in some sense. This will be clear from the examples below.

A definitor, as I think of it, would not have to be formated, but it probably would be. By "formated" I mean that certain positions (or groups of positions) in the definitor would be dedicated to terms of some particular type or source. Such formating can be used to provide additional semantic information or to introduce a syntactic structure into the definitor itself.

If definitors become widely used, whether for document retrieval or for other applications where semantic factors have to be automated, their design and construction will undoubtedly become a subject for study in its own right. In general, I would expect that each application would require a different technique. In any event, the samples I give here are not meant as models for definitors. In fact, only the last of these, the one for OAK, even begins to indicate the concept of a definitor on which most of my comments in this paper are based. The first four samples are discussed only because they are representative of the definitors used in getting the test results reported later in this paper.

The first two are taken from the dictionary used in the first of two test runs on a computer:

DETECTION = DETECTION + GDDF +
SENSORS + SEARCH
RADAR SIGNALS = RADAR SIGNALS + ELT +
ELECTROMAGNETIC + WAVES +
DETECTION

In these the + sign is used (as it is in the COMIT programming language) to separate constituents. "GDDF" is an abbreviation for "GENERAL DETECTION AND DIRECTION FINDING" and "ELT" is an abbreviation for "ELECTRONICS." These are formated according to:

$$T = T + G + A + B$$

where T stands first for the first order descriptor being "defined" and then, on the right of the = sign, for its repetition as one of the second order descriptors comprising the definitor. G stands for a generic term taken from the middle one of three hierarchical levels of the terminology control list used by the MITRE Library. It, therefore, was in a first order generic relation to the descriptor being defined since the actual file descriptors comprised the lowest of the three levels. The A and B stand for additional terms. These were assigned quite freely — in fact mostly while I was sitting at a keypunch — and without referring to the test corpus of documents or to other definitors.

These are about the minimum I would consider as still retaining the flavor of a holistic definition-surrogate in any sense. For while some dictionary definitions do consist of a single synonymous term, this assumes the associative/cognitive powers of a human intellect to therefrom "understand" a meaning. In principle, this could presumably be mechanized by sufficiently iterating the dictionary look-up procedure, but such iterations are not immediately contemplated here. In these minimal definitors the principal — and only systematic — association-making power is provided by the generic terms. DOPPLER SYSTEMS, another first order descriptor used in the test corpus, was classified by the terminology control list in the same group as DETECTION, therefore would have GDDF for its generic definitor constituent, as did DETECTION. Thus, the computer is enabled to "know" that there is some relation between DOPPLER SYSTEMS and DETECTION in terms of some aspect of their meanings *to humans*. Not only is the information that *some* relation exists between these two terms thus made available to the computer, but in this particular example additional information as to the *kind* of relation is gratuitously available because of the formating; i.e., that they both stand in the same specific/generic relation to some term (which can be identified if need be).

Far more information than just exemplified can be provided by definitor systems that are properly designed, implemented, and utilized. Before leaving the sample definitors above, one further instance can be noted. In this case, a nonsystematic relation between DETECTION and RADAR SIGNALS is made available to the computer by virtue of the appearance of DETECTION as the fourth definitor constituent for RADAR SIGNALS. Now, it was my intention — and to the best of my knowledge it was carried out — to assign the last two definitor constituents in this particular set quite freely (not, however, amounting to what a psychologist would call "free association") especially making it a point *not* to refer either to the test documents or to the other definitors. It can, therefore, be considered that the assignment of DETECTION to RADAR SIGNALS, hence the relation made available to the computer by this assignment, was fortuitous. I feel that the appearance of relations in this manner is one of the virtues, rather than one of the vices, of definitors, and that part of whatever power they give the computer to emulate the human's intermediacy will derive therefrom.

The second two definitor samples are taken from the ictionary used in the second of the two test runs:

DETECTION = DETECTION + GENERAL DETECTORS AND DIRECTION FINDING + DETECTORS + SENSORS

RADAR EQUIPMENT = RADAR EQUIPMENT + RADAR AND RADIO DETECTION + SENSORS + ELECTRONIC EQUIPMENT

'he format for these is:

$$T = T + G + A + B(+C)(+D)(+E)$$

'his is the same as before except that here up to three dditional free terms were optional, as indicated by the erms in parentheses. As before, definitors were contructed without reference between themselves or to the ocuments. However, the participation of a librarian ısured more care in their construction. This undoubtedly lso did introduce some knowledge of the nature of the ontents of the test file, but this was unintentional. (I on't mean to imply that, normally, such knowledge hould not be intentionally introduced — here, the intenion was to reduce the number of parameters that would ave to be outguessed in the appraisal, preferably by rring on the side of disadvantage to performance.) The wo factors, more terms and greater care, did appear to how a definite increase in power. SENSORS relates the wo above nonsystematically. Formalizing definitors long such lines as the next example would make such elating less a matter of chance.

As a final sample, the following definitor was not used ı the work reported in this paper, but is taken from the lictionary I constructed for a follow-on study: (2)

OAK = OAK + FOUR + 4III + 4IIIC + 410 + 410.51 + HARDWOOD + TREE + BROADLEAF + LARGE + ACORN + NUT + NOUN + CONCRETE + FAGACEOUS + — + — + DECIDUOUS

'he format for this one can be represented as:

$$T = T + R1 + R2 + R3 + R4 + R5 + F1 + F2 + D1 + D2 + D3 + D4 + G1 + G2 + M1 + M2 + M3 + M4$$

vhere T is the same as before, the R terms are from Roget's *Thesaurus*, the F's from *The Golden Encyclopedia*, a children's book, the D's from Webster's *New Collegiate Dictionary,* the G's are parts of speech, and he M's are for free terms. I had hoped to take addiional constituents from a faceted classification scheme ut time did not permit. This definitor is shown here nly in the interest of bettering communications: somehing more like it is what I really have in mind when alking about definitors unless otherwise indicated.

4. The Pseudometric

To simply replace file descriptors by definitors similar o the last one shown would be disastrous in that it would result in a tremendous number of documents of little or no relevance being retrieved. In trying to escape the devil of paucity and sterility of word associations based on first-order descriptors we would drown in the deep blue sea of too many and too tenuous word associations based on second-order descriptors. Such an increase in "false drops" has no doubt discouraged many attempts to increase the richness of word associations in the retrieval process. Lancaster and Mills stated the generally held belief, "Devices such as confounding of word forms or generic searching may, by enlarging the classes to be searched, improve recall. But they will do this only at the expense of relevance." (3)

Fortunately, quantifying and normalizing functions are available which simply thrive on such multiplicities of associations. The kind of functions I have in mind are, in essence, statistical "mechanisms." As such, the larger the "sample" they have to work from, the better they work. In fact, I suspect that the main reason for their not having been much more effective to date is just that it has heretofore been impracticable to "feed" them with large enough "samples." This need can be satisfied by the definitor in an intuitively meaningful way. This, in turn, enables the quantifying and normalizing function to return the favor by satisfying the need created by the definitor for (a) a means to control the quantity of items retrieved and (b) a means to ensure that, whatever quantity is chosen, it will include the most relevant items, i.e., a means of ordering the items according to relevance to the retrieval request.

The particular function I use is the "pseudometric"[1] introduced and discussed by Rial (4), though such functions as those of Stiles, Maron, and others discussed by Hayes (5) may be able to be used. This calculates a "distance" in a "concept space" between one string of descriptors, say those expressing a query or request to the information store, and another string of descriptors, say those of any of the documents in the store. The formula is simply

$$F = 1 - \frac{|A \cap B|}{|A \cup B|}$$

where F stands for the "pseudometric distance," A and B represent the two descriptor sets, "$\cap$" is the symbol used in logic for the "meet" or "intersection" of two sets, "$\cup$" is the symbol for the "union" of two sets, and the pair of vertical lines is used to denote that the value expressed is the cardinality (number of things) and not the actual indicated things themselves. Thus, the numerator of the fraction part is the number of descriptors the two strings have in common and the denominator is the total number of different descriptors in the two strings.

[1] The reason that this is referred to as a "pseudometric" rather than a "metric" is that while it meets two of the three requirements for a metric (the symmetry requirement, $\delta(x, y) = \delta(y, x)$ and the triangle inequality, $\delta(x,y) + \delta(y,z) \geq \delta(x,z)$) the zero-distance requirement for a "metric" ($\delta(x, y) = 0$ if and only if $x = y$) is relaxed to allow $\delta(x, y) = 0$ even if $x \neq y$.

All "distances" fall into a range of zero to one regardless of the number of descriptors in either string, thus providing normalization. Subtracting the fraction part from unity has the incidental but happy result of making the calculated results compatible with our intuitive feeling for distances. That is, a distance of "zero" shows maximum conceptual closeness (when the two strings are identical) and a distance of "one" shows complete irrelevance (when the two strings are entirely different).

The way the pseudometric works — and the way it complements the definitor — will be clear from the sample calculations to be given. First, a few comments about my choice of this particular function: most important is that it is the simplest function I know of which satisfies the needs introduced by the definitor. Also it is intuitively comfortable to work with. It does, however, have weaknesses. The most noteworthy of these may be illustrated by the following case:

Documents	*Descriptors*
A	j
B	j, k, l, m

then

$$\hat{F} = 1 - \frac{1}{4} = .75$$

Since the subject of document A is indicated as being completely encompassed by the subject matter of document B, to say they are so far apart seems questionable. Yet, on the other hand the added subject matter of B makes it unlike A. Would we want to consider using some function that would show A as being closer to B than B is to A? Or would we be able to develop a correction factor for such cases, where the string lengths are different? I don't know, but I do believe that if we remember that the problem posed here is a psycholinguistic one rather than an arithmetical or logical one, we'll be able to handle it even if different means are needed in different environments.

• 5. The Procedure

To help follow the calculations and discussions in the ensuing sections, the operational retrieval procedure I am proposing is outlined here in its barest essentials. I assume that we are talking about fully automated retrieval. For simplicity at the moment I assume that queries to the system are expressed in terms of a string of descriptors for each of which there is a definitor in the dictionary of the same kind as those used to define the document descriptors used in the system.

First, each query descriptor would be replaced by its definitor by means of an automatic dictionary look-up. Then, in turn, each document in the file would have its descriptors replaced by their definitors and a pseudometric distance between it and the query would be calculated on the basis of the two strings of second-order descriptors. Then each pseudometric distance so calculated would be compared to a preset cut-off value (which I assume could be changed at will between runs). All documents whose pseudometric distance from the query was greater than the cut-off value would be discarded. Those remaining would be ordered according to their distances, those with the least distance first, and presented, together with the calculated distances, as the retrieval result.

The foregoing is intended only as an outline. By "presenting" a "document" in the output I mean anything from an identifying symbol to the document itself — it's the selection process we're concerned with. "Each document" could be "each document that had survived some preselection routine, such as taking the query definitor constituents through a concordance." The cut-off value could as well be applied to each pseudometric distance as soon as it was calculated. And obviously, if it was desired to buy time by paying storage space, the substitution of definitors for the document descriptors could be done when the document (representation) was stored. Of course, then the file (rather than just one entry in the dictionary) would have to be updated whenever a definitor was changed in any way.

• 6. Feasibility Demonstration

As a feasibility test of the proposed procedure, I wrote a computer program in the COMIT programming language[2] and used it to make one test run of five queries against a test corpus of 41 document representations and another test run of three different queries against a different test corpus of 32 document representations. The way in which this program works is illustrated in some detail to give a better picture of the nature of the proposal. The fact that the program exists and ran successfully merely proves that it is indeed feasible to construct a program which will implement the proposed procedure and that it will in fact operate on a computer as planned. It says nothing about the validity or the worth of the procedure itself.

To provide a basis with which to compare the retrieval results achieved by using the second-order descriptors, the program first calculates the pseudometric distance between a query and a document using only the regular file (first order) descriptors. It should be realized that this would not be done in an operational environment; it was done simply for evaluation purposes. For ease of reference we will call the distance so calculated the "first-order distance" to distinguish it from the "second-order distance" resulting from use of the second-order descriptors.

Figure 1 shows how the first-order distance was calculated. The query and document used in the figure are taken from the second test run, so the calculation shown is one of those actually done by the computer. The query

[2] Developed by V. H. Yngve at MIT (6).

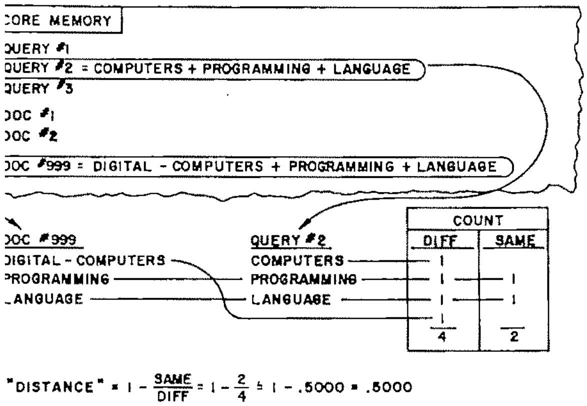

:ORE MEMORY
QUERY #1
QUERY #2 = COMPUTERS + PROGRAMMING + LANGUAGE
QUERY #3
DOC #1
DOC #2
DOC #999 = DIGITAL - COMPUTERS + PROGRAMMING + LANGUAGE
DOC #999
DIGITAL - COMPUTERS
PROGRAMMING
LANGUAGE
QUERY #2
COMPUTERS
PROGRAMMING
LANGUAGE
COUNT
DIFF
SAME
1
1
1
1
4
1
1
2
"DISTANCE" = 1 - SAME/DIFF = 1 - 2/4 = 1 - .5000 = .5000

Fig. 1. Calculation of first-order distance

scriptors are brought in from storage and counted to art off the count of different descriptors. Then the cument descriptors are brought in from storage one a time and checked against those of the query. Here, e first one, DIGITAL COMPUTERS does not find a atch so a count is added to the count of different scriptors. (To find a match, of course, the computer ust find a perfect character-by-character coincidence tween the two terms.) The next, PROGRAMMING, ıds a match so a count of one is recorded to start the unt of "same" descriptors. The last, LANGUAGE, so finds a match adding another "same" count. The st-order pseudometric distance is therefore

$$F = 1 - \frac{\text{No. same}}{\text{No. different}} = 1 - \frac{2}{4} = 1 - .5000 = .5000$$

his is recorded in temporary storage and the program oceeds to its principal task, the calculation of the cond-order distance.

Figure 2 shows schematically how the two strings of cond-order descriptors are formed by replacing each scriptor in the query and in the document with its finitor. (Actually once the string of second-order deriptors for the query has been formed for use against the first document it is saved for use against succeeding documents, eliminating pointless repetitions.)

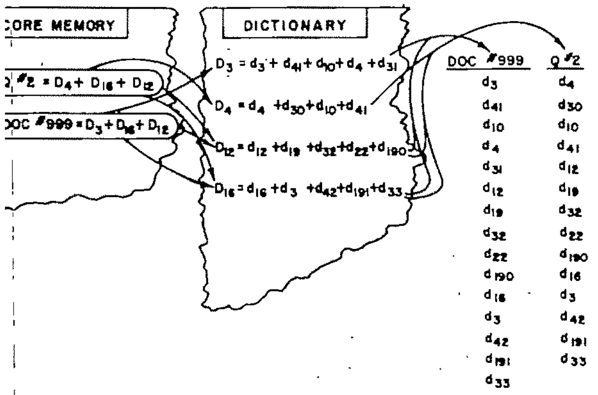

:ORE MEMORY
DICTIONARY
Q #2 = D4 + D16 + D12
DOC #999 = D3 + D16 + D12
D3 = d3 + d41 + d10 + d4 + d31
D4 = d4 + d30 + d10 + d41
D12 = d12 + d19 + d32 + d22 + d190
D16 = d16 + d3 + d42 + d191 + d33
DOC #999
d3
d41
d10
d4
d31
d12
d19
d32
d22
d190
d16
d3
d42
d191
d33
Q #2
d4
d30
d10
d41
d12
d19
d32
d22
d190
d16
d3
d42
d191
d33

Fig. 2. Formation of second-order descriptors

Figure 3 shows how the second-order distance was calculated. As before, the count of "different descriptors" is started off by counting all the query descriptors, then a one-by-one check of the document descriptors against the query descriptors is made.

To clarify what is happening, the first term in each definitor (which, it will be remembered was just a repetition of the term being defined) is underlined. Note that COMPUTERS appears twice in the query's string, once as the repetition term in the definitor for COMPUTERS and then as the second term in the definitor for PROGRAMMING (since it was the term generic to PROGRAMMING in the terminology control list used). As the program works at present, these two occurrences of COMPUTERS are each counted as an additional "different descriptor" in the first count of the query descriptors. Furthermore, note that COMPUTER also appears twice in the document's string. Each of these, when being checked against the query descriptors finds a match immediately and the count of "same" descriptors is increased by one for each of them. The check stops, for each term, when a match is found so that the further match possibility is not found. The matter of how such repetitions in either string should be handled needs further study, especially since it would seem intuitively that just such repetitions would be particularly significant as indications of conceptual closeness. The way they were handled here tends to yield a smaller distance in such cases, as seems proper, but if the second occurrence of COMPUTER in the query string had not appeared there but instead had been a third occurrence in the document string the count would appear to indicate a logical absurdity, i.e., that the set intersection was greater than the set union. The distance would have come out

$$1 - \frac{15}{14} \text{ or } 1 - 1.0714 \text{ or } -.0714$$

Perhaps we should accept such a result as a signal of "extra closeness!"

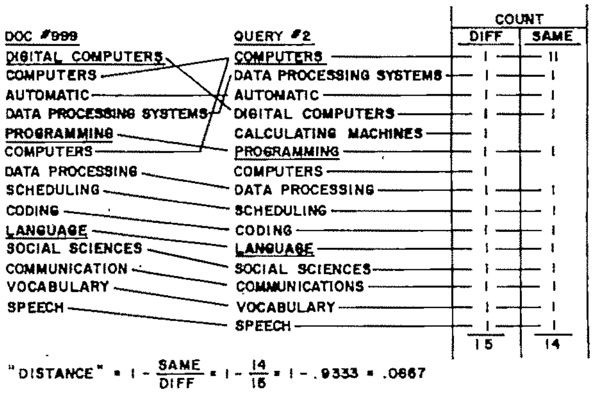


Fig. 3. Calculation of second-order distance

By so defining the quantifying and normalizing function as to permit it to range into the negative real numbers as well as the non-negative real numbers, with ordering still done algebraically (e.g., $-3 < -2$), it seems possible that greater latitude is made available for comparisons of close "distances" occurring in some such manner as that just indicated above. Such latitude just might make an iteration procedure more worthwhile than it would be otherwise, for example.

To return to the computation, the second-order distance for this query-document pair came out to be .0667, much more as, intuitively, it should be. The program then records this in temporary storage and proceeds to calculate the two distances between the same query and the next document in the file. The results of applying the cut-off value to the resulting first- and second-order distance lists, and of then ordering, labeling, and printing out those left is shown in the next section.

• 7. Results

The results obtained from the two test runs are presented in Tables 1 through 8. The ordered lists and associated pseudometric distances are copied directly from the computer print-out. The figures in the single digit columns give the relevance of each document to the query as judged by a professional librarian. A "0" indicates "complete relevance" and signifies that, to the satisfaction of the judge, the document would have given a sufficient answer to the query all by itself. A "1" represents nearly complete relevance in the sense that the judge considered that the document almost — but not quite — answers the query by itself. A "2" indicates that the document was still considered fairly relevant but not quite so much so, and so on. A dash indicates that the document was not considered at all relevant to the particular "need" represented by the query. Brackets have been added to show those documents calculated as being at the same distance as a group. In such cases, the order within the group is simply the order in which the computer came to them, i.e., their original order in the file.

A discussion of Query 7 (see Table 7), which is already somewhat familiar from the preceding section, will clarify the meaning of these results. When the regular file descriptors were used, 7 documents survived the application of the cut-off value and were ordered as shown. When the second-order descriptors were used, 5 additional documents were retrieved and the 12 were ordered as shown. One which was judged not relevant, at the bottom, just squeaked by the cut-off value which was set to drop anything above .9990.[3] Notice how Document No. 33944, used in the illustrative example in the last section, was moved from second place on the first-order list to an unequivocal, and intuitively correct, position in first place on the second-order list. Note also the greater over-all ordering power shown by the second-order list compared with that shown by the first-order list.

The procedure had its only clear-cut failure with Query 5 where it not only did not uncover any additional relevant documents — or improve on the ordering of the first-order search — but dug up seven "trash" items, documents that were judged to be totally irrelevant as

TABLE 1. Query 1—Documents on cartography, specifically those relating to methods of producing maps automatically by digital means.

First Order			Second Order		
Doc. No.	Distance	Judge	Doc. No.	Distance	Judge
[22088	.8335	2	22088	.4000	2
20339]	.8335	4	15101	.5712	0
15101	.8574	0	15392	.6660	1
[15392	.8750	1	20339	.6664	4
22079]	.8750	3	22079	.7364	3
			23145	.8568	5
			22078	.9163	6
			[23121	.9338	—
			23190]	.9338	2
			23198	.9570	—

TABLE 2. Query 2—All documents on ablation and reentry vehicles.

First Order			Second Order		
Doc. No.	Distance	Judge	Doc. No.	Distance	Judge
[10005	.6666	1	[10005	.3750	1
13474]	.6666	2	13474]	.3750	2
22082	.7500	0	22082	.5831	0
			08318	.8671	3
			[20340	.9163	4
			23169	.9163	5
			20341]	.9163	6

TABLE 3. Query 3—All documents on harmful atmospheres.

First Order			Second Order		
Doc. No.	Distance	Judge	Doc. No.	Distance	Judge
23181	.7500	0	23181	.4545	0
			23183	.6921	0
			23182	.7145	1

[3] Actually, this document had no descriptors in common with the query, so should have been distanced at 1.0000, thus dropped by the cut-off value as set. The reason it wasn't is as follows: Because of the difficulty of doing division in COMIT (which otherwise is excellent for this application) I converted the pseudometric as shown:

$$F = 1 - \frac{|A \cap B|}{|A \cup B|} \Rightarrow \frac{|A \cup B|}{|A \cup B|} - \frac{|A \cap B|}{|A \cup B|} = (|A \cup B| - |A \cap B|)\left(\frac{1}{|A \cup B|}\right)$$

and programmed a table look-up of the reciprocal of $|A \cup B|$ which was then added to itself a number of times equal to $(|A \cup B| - |A \cap B|)$. This resulted in accumulating any round-off error in the reciprocal, as in this case, where $|A \cup B|$ was counted as 27 and the reciprocal of 27 was given in the table as .0870, which, added to itself 27 times, gave .9990.

TABLE 4. Query 4—Automatic language abstracting.

First Order			Second Order		
Doc. No.	Distance	Judge	Doc. No.	Distance	Judge
22079	.4000	0	22079	.2940	0
20339	.5000	0	20339	.3845	0
22088	.8000	3	22088	.6003	3
15101	.8335	4	15392	.6825	2
15392	.8574	2	15101	.7500	4
			23105	.9135	5
			22078	.9408	7
			23188	.9500	—
			23121	.9500	—
			23145	.9500	8
			23159	.9591	—
			22081	.9591	—
			23198	.9639	—

far as answering the query was concerned. However, a cut-off of .9300 would have eliminated all of these false drops.

The procedure had a clear-cut success with Query 3 where it uncovered one "direct hit" that had not been found by the conventional search and another highly relevant document and properly ordered them. The apparently remarkable ordering power shown by Query 7 was also shown by Query 6. The ordering in Query 8, while not as spectacular as the preceding two, nevertheless would seem to be good enough to perform a useful function. Comparing the results of the second run (6, 7, and 8) with the first (1 through 5), it is easy to imagine that beneficial effects of the slightly longer and slightly more carefully constructed definitors are manifesting themselves in terms of increased ordering power.

In general, the retrieval based on the second-order descriptors produced more documents in every case — as would be expected with the cut-off set only to eliminate documents that had no descriptors in common with the query. The significant question is: Was there evidence of sufficient improvement in the ordering power to promise that when the cut-off value is used as a control to limit the number of documents retrieved, the most relevant ones would be retained?

TABLE 5. Query 5—Display systems and aerial photographic technique.

First Order			Second Order		
Doc. No.	Distance	Judge	Doc. No.	Distance	Judge
23188	.0000	0	23188	.0000	0
23190	.6666	0	22081	.5712	0
22081	.7500	0	23190	.6664	0
			22077	.9338	—
			20339	.9338	—
			22088	.9338	—
			22552	.9468	—
			23105	.9468	—
			15392	.9570	—
			22079	.9570	—

TABLE 6. Query 6—All documents on weapons effectiveness in military operations.

First Order			Second Order		
Doc. No.	Distance	Judge	Doc. No.	Distance	Judge
33992	.3333	0	33992	.2668	0
33903	.6000	4	33932	.4704	1
33932	.7500	1	33906	.4704	2
33906	.7500	2	33901	.6312	3
33901	.7500	3	33903	.6400	4
33943	.8000	8	33916	.7600	5
33920	.8000	7	33915	.7600	6
33916	.8000	5	33920	.7700	7
33915	.8000	6	33943	.8400	8
			33933	.8645	9
			33975	.8757	10
			33961	.9591	—
			33960	.9591	—

Insofar as these test results are indicative, the answer to this question would seem to be yes. For one thing, of an over-all total of 42 documents retrieved by the second-order searches over and above those that had been retrieved by the first-order searches, all but 17 were judged relevant to the query to some extent. None of these 17 "completely irrelevant" documents would have been retrieved if the cut-off value had been moved by as little as .9990 to .9300. The cost, in terms of relevant documents discarded with them, would have been four; these had relevancy judgments of 2, 7, 8, and 9 (in Queries 1, 4, and 8). For another thing, as remarked before, several of the individual query results do show a definite increase in ordering power while none shows a decrease.

The fact is, however, that generalization from these results is not warranted. Even if we knew what the

TABLE 7. Query 7—All documents relating to computers and programming language.

First Order			Second Order		
Doc. No.	Distance	Judge	Doc. No.	Distance	Judge
33967	.3333	1	33944	.0667	0
33944	.5000	0	33967	.3335	1
33924	.6666	3	33941	.4002	2
33923	.6666	4	33924	.6670	3
33941	.7500	2	33923	.6670	4
33925	.7500	5	33922	.6838	4
33922	.7500	4	33925	.7364	5
			33965	.7500	6
			33968	.7890	7
			33912	.7890	7
			33943	.9250	8
			33975	.9990	—

Table 8. Query 8—All documents on detection of satellites and guided missiles.

First Order			Second Order		
Doc. No.	Distance	Judge	Doc. No.	Distance	Judge
33920	.8335	2	22920	.7395	2
33918	.8335	1	33917	.7770	0
33917	.8335	0	33918	.8092	1
			33912	.8880	4
			33911	.8942	3
			33935	.9000	6
			33975	.9100	7
			33953	.9240	8
			33919	.9250	10
			33954	.9639	9

important parameters were in such an interplay of documentation, linguistics, statistics, and psychology as we have here, we couldn't say how they interrelate. Consequently, it cannot be guaranteed that the results just discussed could not have been produced by a fortuitous combination of atypically advantageous factors. Therefore, these results must be considered indicative rather than conclusive. This is not to say that they cannot be considered generally as quite encouraging. I so consider them and hope to soon have an opportunity to plan and conduct better designed and more extensive tests.

• 8. Possibilities

In this section I point out some of the ways in which the definitor-pseudometric combination, by exploiting relations inherent in natural language terms, would seem to offer increased efficacy in the retrieval of information from existing descriptor-indexed files without requiring that they be re-indexed. I believe that, as with the basic proposal itself, the validity in principle and the feasibility of computer implementation of each of these is fairly self-evident. Again, as with the basic proposal, their worth under various sets of circumstances remains to be explored. In other words, no more is being claimed for them than for the basic proposal, although, for simplicity of presentation, repetitions of such qualifications and disclaimers will be omitted.

First, and perhaps most important, is that new terminology can be introduced. For example, the term "laser" along with a definitor for it can be put in the dictionary even though the file may have been generated and indexed before lasers were invented. Then if "laser" is used in a query it will be interpreted into other concepts which may find relevant material in the file. "Laser" can also be added to appropriate terms of a file as a definitor constituent. For example, such insertion would serve to create or reenforce a link between "coherent radiation" and "stimulated emission" which may not have been considered important when the file was first set up.

Biases may be introduced deliberately through the mediation of the definitors. A mining company, a watchmaker, and a physicist would presumably think of "ruby" in quite different ways and want correspondingly different associations with other terms to be made for retrieval purposes.

If they are known, biases introduced by an indexer can be corrected or compensated for. Suppose one file was known to index documents on "failure" under "reliability," whereas another did not. To some extent different definitor dictionaries could help correct for this difference. In this connection it seems possible that statistical word association techniques might be able to assist by detecting, and perhaps even quantifying, such biases.

Queries and document descriptors can be put through different dictionaries, each designed for specific effects. This would be particularly useful in the sometimes crucial and often frustrating searches by a person of one discipline in files created by, or for, persons of other disciplines. This technique could be extended, if found worthwhile. For example, a query from a member of the physics department could be put through a different dictionary than one from a member of the electrical engineering department, etc. In principle this could be extended to individuals.

The four just discussed are more a system design consideration than otherwise. If they were not available, no one would wait for them; if they were part of the system, there might not be any alternative to using them. The next four are more the sort of thing one might think of as being used on-line as part of the "dialog" between the man and the machine that time-sharing is soon to bring into reality. The first of these has already been mentioned: the on-line adjustment of cut-off values according to retrieval returns. The querier would enter his query using an initial cut-off value determined by experience in some way. This initial setting could be different for different circumstances if experience showed such to be appropriate. It could be determined automatically for each query if desired. An attractive and easily implemented option here could have the computer, after making its search, inform the querier how many documents he was about to get before actually giving them (or their representations) to him. This would enable him to vary the cut-off to reduce (or increase) the number received before accepting the output results.

Various combinations of weightings, threshold values, go-no-go requirements, and logic specifications could be applied to various specified definitor constituents or combinations of them to control the search process. It would be presumptuous, until the applicability of the basic proposal has been more thoroughly explored, to discuss such possibilities in detail or at length. To clarify the point however: assuming, for illustrative purposes only, definitors formated like the one given earlier for OAK, the querier might want to require that a document to be selected at all must have one or more matches with, say, 4III in the third definitor-constituent position. This would insure that each of the documents retrieved would have something to do with organic matter (as such is considered by Roget at least!). As techniques are developed in the interface area between artificial intelligence and linguistics, such tactics as this might be made automatically adaptable to various circumstances or parameters.

Selection subroutines can be installed to cause the dictionary to look to the program as if it consisted only of some specified subset of the constituents in each definitor. Thus, if the querier wanted a connotative, browsy search he might want to use only definitor-constituents which had been derived from Roget. If he wanted a more denotative, definitive search he might want to use only those he knew to have been taken from a technical encyclopedia. Single or double concordances can be combined effectively with such selection subroutines if over all system design considerations so indicate. Figure 4 shows how this might work. The selection subroutine would not merely affect the calculation of the pseudometric distance (see arrows from "Dictionary" down to "Doc 20" and "Q"), but would in the first place have affected the list of definitor constituents used for entry to the "First Concordance" (which for each second-order descriptor would list the first-order descriptor in whose definitor it had appeared). Such concordances, to be practicable, would have to be automatically prepared, updated, and purged from file entries and changes thereto.

Last, but by no means least, the definitor provides a way to get classification back into the picture. By dedicating a number of positions in a formated definitor to constituents from various levels or facets of a classification, matches will be generated according to the structure of the classification scheme. This introduces a self-weighting mechanism in the case of hierarchical classifications because two terms falling into the same class at some level would result in matches also being generated on each higher level of that classification scheme which is used in the definitor format. The way this works can be seen with reference to the way the definitor-constituents from Roget were formated in the illustrative definitor shown for OAK. A definitor constructed in the same format for MAPLE would provide not one but five matches with OAK since "maple" is listed by Roget in the same most specific class, 410.51, as is "oak," and so would also have the same definitor-constituents (which in this case are codes designating Roget classes rather than natural language terms) for all four generic levels.

• Conclusions

The proposed procedure appears to rest on a sound rational basis. Results obtained in two demonstration runs are confirmatory and encouraging but are not conclusive. Therefore, adequately designed and carefully implemented tests to confirm or deny the general validity of the procedure should be conducted. Upon receipt of affirmative results, parametric studies of alternatives,

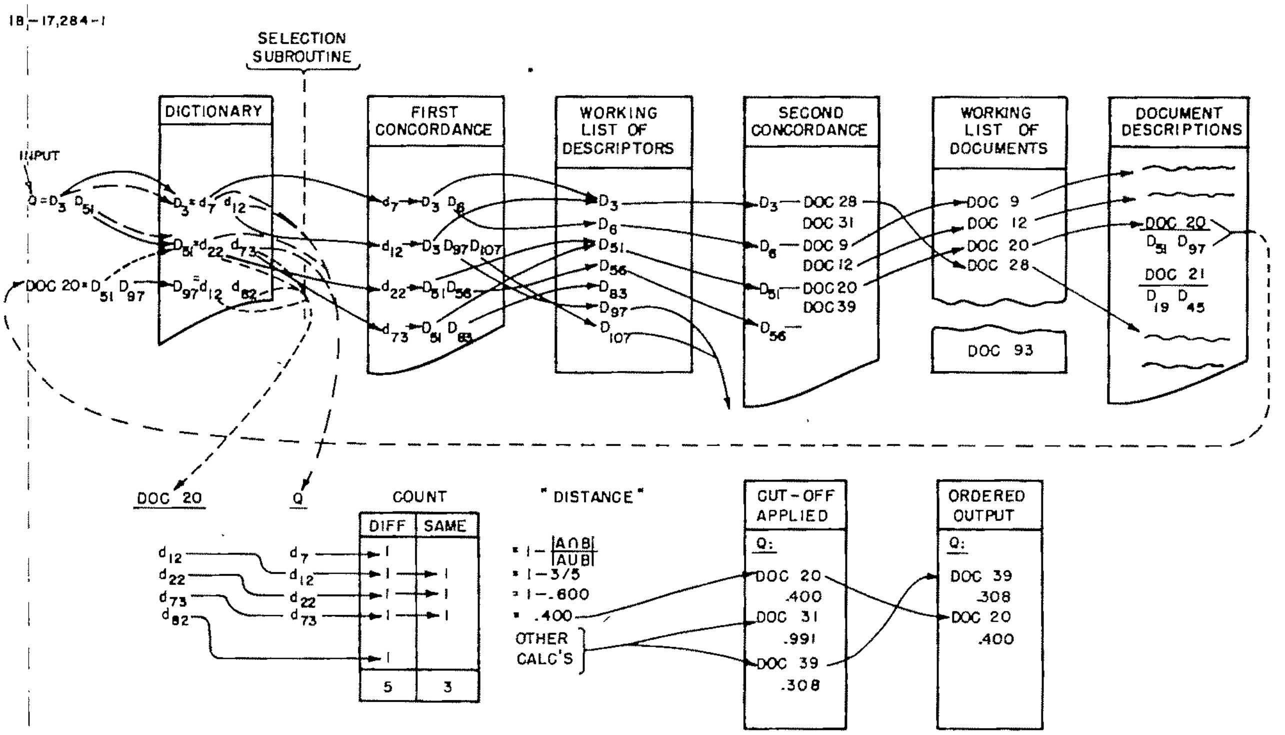


Fig. 4. Procedure incorporating concordances and selection subroutine

tradeoffs, costs, worth, and so forth should be made to provide the kind of information that would be needed by system designers interested in applying it.

Acknowledgments

The germinal idea for this work occurred to me about three years ago, but I had no opportunity to develop it until some two years later when I was allowed to use it as the basis for an MIT term project. This included limited access to an IBM 7094 computer at the MIT Computation Center, enabling the demonstration runs reported herein to be made. (The limitations to this access, and the fact that all of this effort was in addition to full-time work, are the reasons that a more extensive test was not possible.) For providing this opportunity and for his subsequent encouragement, I should like to thank Professor V. H. Yngve for whose course, "Language, Symbolic Processes and Computers," the term project was done.

The test corpus and queries used in the demonstration runs were provided by Mr. E. X. Casey of the MITRE Corporation Library. He also acted as judge, providing the ratings of relevance of documents to queries shown in the "Results" section. For this essential assistance, I want to express my sincere thanks to Mr. Casey.

This paper was presented in a condensed form on October 14, 1965 at the 31st Congress of the International Federation for Documentation, held in Washington, D. C. I wish to express my appreciation to the MITRE Corporation for granting the time and facilities needed to prepare that oral presentation and this written version, as well as for the sponsorship of the MITRE Corporation Educational Assistance Program which made it possible for me to take Professor Yngve's course in the first place.

References

1. Vickery, B. C., *On Retrieval System Theory,* Butterworth's, London, 1961.
2. Libbey, M. A., The Representation of Meaning to Computers, *Proceedings of the American Documentation Institute,* 1966 Annual Meeting, Vol. 3, 1966.
3. Lancaster, F. W., and J. Mills, Testing Indexes and Index Language Devices: The ASLIB Cranfield Project, *American Documentation,* 15:4 (1964).
4. Rial, J. F., *A Pseudo-metric for Document Retrieval Systems,* The MITRE Corporation, Working Paper W-4595.
5. Hayes, R. M., Mathematical Models for Information Retrieval in *Natural Language and the Computer,* Paul L. Garvin, ed., McGraw-Hill, New York, 1963.
6. M.I.T., The Research Laboratory of Electronics and the Computation Center, *An Introduction to COMIT Programming* and *COMIT Programmers Reference Manual,* The M.I.T. Press, November 1963.

Dictionary Buildup and Stability of Word Frequency in a Specialized Medical Area*

This is a report of word usage in radiological (x-ray) patient records as found in a 5% sample of the annual case load at UAMC including 100,000 words. Records were taken exactly as dictated. The study is part of an effort to develop an IR system for patient data. The system "autocodes" (automatically stores) the physician's dictated findings and diagnoses in such a fashion that they can be retrieved again automatically.

Some of our findings approximate results reported in the literature. For example, the rate of introduction of new different words levels off to about 2,500 words when 40,000 to 50,000 words of text have been analyzed. However, unclassified words continue to occur at a significant level of almost 2% at the 100,000 word level, with a 1% noise level.

Attempts to establish the rank order of words beyond the first several hundred have failed because about 70% of the words appear to occur with such a low relative frequency (no more than one time in 10,000). Thus, establishing files by rank order appears impractical, even though filter lists (discard words) by rank groups (words with nearly the same relative frequency) are quite practical.

Additional data are presented and design implications are discussed.

JOHN M. LONG, HOWARD J. BARNHARD, AND GERTRUDE C. LEVY†

University of Arkansas Medical Center
Little Rock, Arkansas

1. Introduction

This report covers a word analysis of the running text of radiologists' dictation. We have analyzed the first 100,000 words in 3,321 radiological (x-ray) records comprising 170,000 words. The study is a part of a major effort which has been underway for several years at the University of Arkansas Medical Center, and is devoted to the development of an information storage and retrieval system designed to handle radiological patient data.

The over-all system is explained in an earlier paper which appears in the April 1966 issue of *American Journal of Roentgenology* (1). The system that we are developing is one in which the physician's dictated findings and diagnoses are automatically stored in a key word type index in such a fashion that it can be retrieved again automatically by use of key words, or in a number of additional ways.

The term "key word" as used here stands for a subset of the English language consisting of words which are considered to be important for some specialized area, in this case, radiology. These are words that radiologists consider important or those that have special technical meaning to radiologists. This subset probably will include a maximum of 2,500 to 3,500 words.

In addition, there will be a nonoverlapping subset of at least 2,500 to 3,500 words which are used frequently in the text of the radiologists' dictation. At the same time, they are words which have little or no information useful to the radiologist. These are called "discard" words.

Words that are not selected for the key word or discard list are called unclassified words. In our system such words are printed out for a specialist to analyze and determine which of three fates the word will have; either it is added to the key word or to the discard list, or it is left as "noise" if it is not used frequently enough and does not have sufficient information to be considered

* Assistance for this research was provided by the University of Arkansas Medical Center Research Computation Laboratory, which is supported by NIH Grant FR 00208–02.

† Respectively, Associate Professor and Head, Biometry Division; Professor and Chairman, Department of Radiology; and Systems Analyst, Research Computation Laboratory.

suitable for either list.[1] Figure 1 presents a Venn diagram of our concept of how the words of the English language will break down in our system.

The principal reason for the studies reported in this paper was to determine the characteristics of our data so that we would have the information needed to properly design our system. A few of the pertinent questions in this regard might be: How much data are needed to start the system going? What are the advantages of a file organized by the rank order of the words by frequency of use over one organized by the alphabet? How difficult is it to determine the rank order of the words? How serious is the "noise" problem? How large must the discard word lists be to reduce the "noise" level to some specified amount such as 1% or .01%? Part of the answers to these questions can be found in the literature as indicated in the next section.

• 2. Literature Review

In 1923, Godfrey Dewey (2) analyzed 100,000 words of running texts of general English language material; his is the most comprehensive study of this type that we know about. He found that "the," "of," "and," "to," "a," and "in" are the six most frequently used words in general English text (see Table 1). Dewey stated, as a rule of thumb, that the 10 most common words would form over 25% of the total words, the 100 most common words would form over 50%, and the 1,000 most common words would form 75% of the total number of words.

TABLE 1. Relative frequency of common words (expressed as a percentage).

Word	This study	Dewey's study
the	9.98	7.5
of	4.32	4.0
and	3.78	3.3
is	3.10	3.3
to	1.19	2.9
a	1.25	2.1
in	1.79	2.1

Andrew D. Booth (3) recently compared Dewey's list with a similarly constructed list in a specialized area. Using his novel approach, if the frequency of use of a particular term in the specialized area differs greatly from its use in general text, it is taken to be a "key word" for this specialized language.

Eugene Schwartz (4, 5) and others point out that a specialized vocabulary will build rather rapidly to about 2,500 words when approximately 40,000 words of running text have been processed.

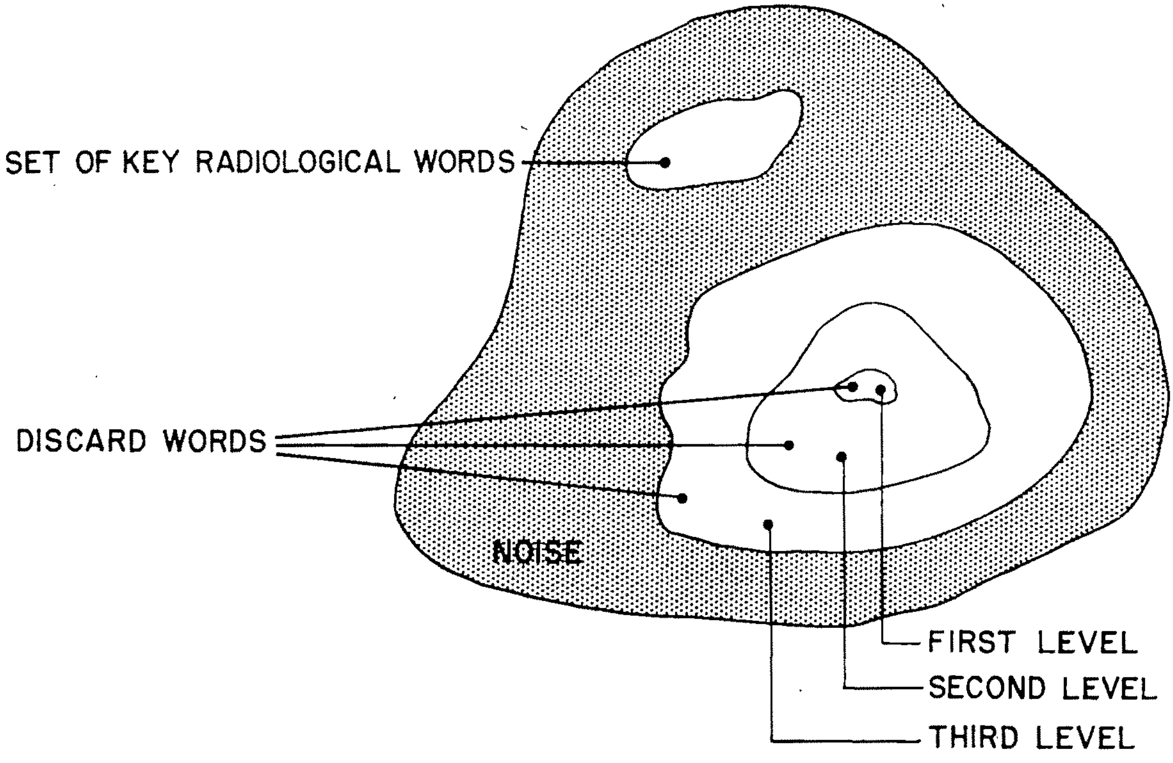


FIG. 1. Set of English words

[1] In our system it appears necessary to classify virtually all words used in order to keep "noise" to a tolerable level.

3. Methods

The raw data for the study are the vocabulary used by radiologists in dictating their findings and diagnoses as found in 3,321 radiology case records. The cases were selected so as to provide a representative sample of the approximately 45,000 cases handled by the UAMC Radiology Department each year.

The reports have been taken exactly as dictated,[2] and transcribed onto punched cards or paper tape for analysis using the IBM 1401 Data Processing System. A number of computer programs were written to provide the analyses reported. The authors will provide additional information about these programs to those interested.

4. Dictionary Buildup and Other Data

We found as did Schwartz (5) that, at about 40,000 words of text, the first 2,500 different words had been introduced and the introduction of new different words had reached a relatively low level. This is shown graphically in Fig. 2. The dotted line, using the left hand scale, shows the rate at which new words are introduced into the system.

Most of the "key" words in a specialized area will have been introduced when approximately 40,000 words of running text have been reviewed, provided the sample is representative. In theory, the rate of introduction of new technical terms should approach zero. However, it is still significant at the 100,000 word level which is as far as our data currently goes. Probably somewhere after the retrieval system has been operating for a long time and the number of key words has probably gone up to 3,500 or more, the rate at which new technical information terms will be introduced will be quite small.

The problem of unclassified words and noise is quite persistent, since unclassified (new) words continued to occur at a significant rate for a long time. This is illustrated in Fig. 2. Refer to the solid line and the right hand scale. Either a highly trained and expensive man, or the computer, must evaluate the unclassified words to separate "noise" from "information" (key words) and "persistent noise" (discard words). Noise can be reduced to as low a level as desired, except that it can never reach absolute zero. The tolerable level of noise varies with the application. We think our system should have considerably less than 1%. In the 40,000 to 100,000 word range unclassified words occur at about 1.8%. Thus, for every 1,000 words of running text introduced, about 18 will be unclassified, with about 2/3 or 12 as "noise" and potential discard words, and 1/3 or 6 as new technical "key" words.

Our system uses a series of discard lists roughly grouped by rank. We refer to these rank groups as discard list filters. The first "filter" eliminates about 30 of the highest frequency words and reduces noise by about 50%. A secondary "filter" of about 2,500 to 3,500 discard words next highest in frequency reduces "noise" to about 1%. We feel this level is still too high. It is not clear whether it would be more efficient to enlarge the secondary "filter" list or to create a tertiary "filter" of discard words to reduce noise to a tolerable level.

We found that the first 10 words represented 31.8% of the total words used; the first 100 words, 64.9%; the first 1,000 words, 94.5%. These results follow Dewey's rule as stated earlier, but our percentages are higher.

Looking at the problem another way, we found that "the" consistently comprised approximately 10% of the words used (actually 9.98%), "of" comprised 4.32%, "and" 3.78%, "is" 3.10%, "to" 1.19%, "a" 1.25%, "in"

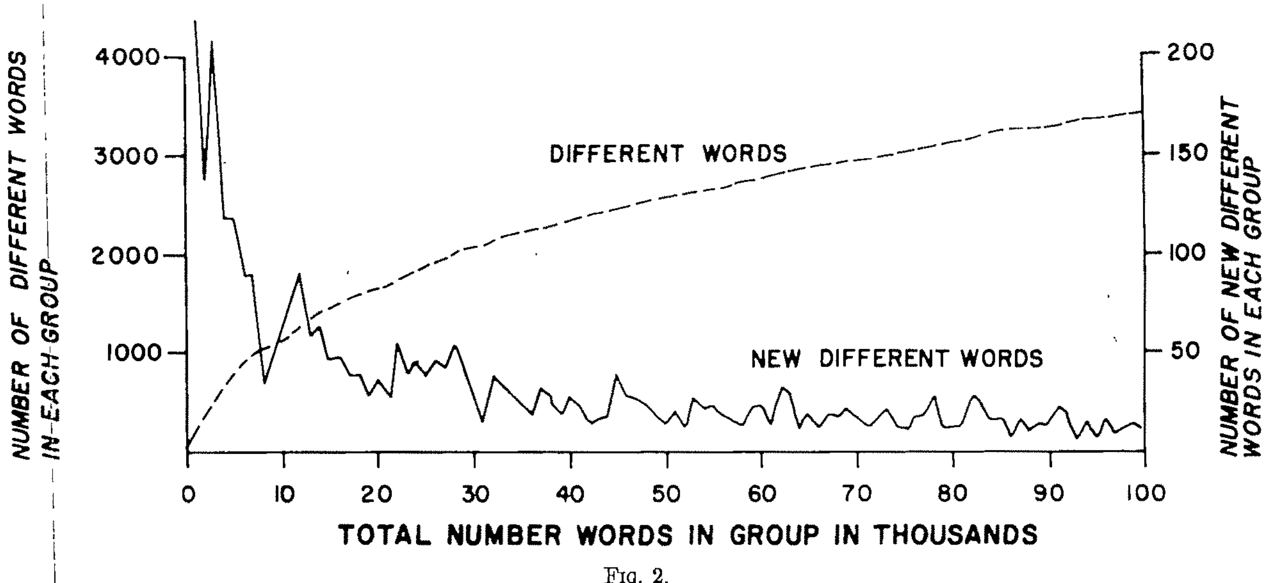


Fig. 2.

[2] The persons transcribing these dictations make mistakes in typing and in spelling, and these were corrected manually. Incidentally, the latter especially is a serious problem and is being studied in depth as another phase of the project.

1.79%. This is summarized in Table 1 with a comparison to Dewey's results. There is a significant difference in the frequency of the use of these words in radiology dictation over that of the general English text as presented earlier. The results suggest a potential flaw in the logic of Booth's (3) proposal.

These findings have definite implications for design. We considered a retrieval system designed so that no key words or discard words were initially determined, that is, you simply start inserting running text of the specialized area and analyzing the words. Initially all words would be unclassified and thus all of them would be printed out. A specialist in the area would determine those words which are considered key words and those which would be considered discard words. The words would be inserted back into the system and the system now would begin to store information about key words and to ignore discard words.

To attempt to build such a system without a preliminary analysis of some text material does not seem to be practical because the number of key words and discard words that would have to be added in the beginning would be at such a rate as to make it rather difficult to handle. Also, it would require a great deal of back updating since when a new "key word" is added it must be coded for all previous cases where it was used before it was selected as a key word. [See the earlier report (1).]

It would be more practical to develop a retrieval system by analyzing approximately 40,000 to 50,000 words of running texts, using the key words so found as a starting list. The other words found with a reasonable frequency would furnish a beginning for the discard word list. Actually, using this method, we compiled a list which is more comprehensive and more suitable for computer coding than the previous lists of terms that have been published by radiologists, such as the *Code for Roentgen Diagnosis Indexing* (6).

• 5. Stability of Word Frequency

The studies of rank stability have implications regarding the method of searching the lists. The idea of searching by rank rather than by the alphabet has logical appeal. However, many questions must first be answered, such as: Can a stable list by rank be achieved, and if so, how long will it take? Once done, will the time saved be worth the effort?

Our results tend to answer this last question in the negative for it is indeed difficult to get rank stability. We found that the ranks of the first several hundred high frequency words are approaching stability when about 80,000 words have been analyzed. The 25 or 30 highest frequency words are rather apparent when only 10,000 or so words have been processed. Figure 3 shows the rate of approach to rank stability for four selected words. The

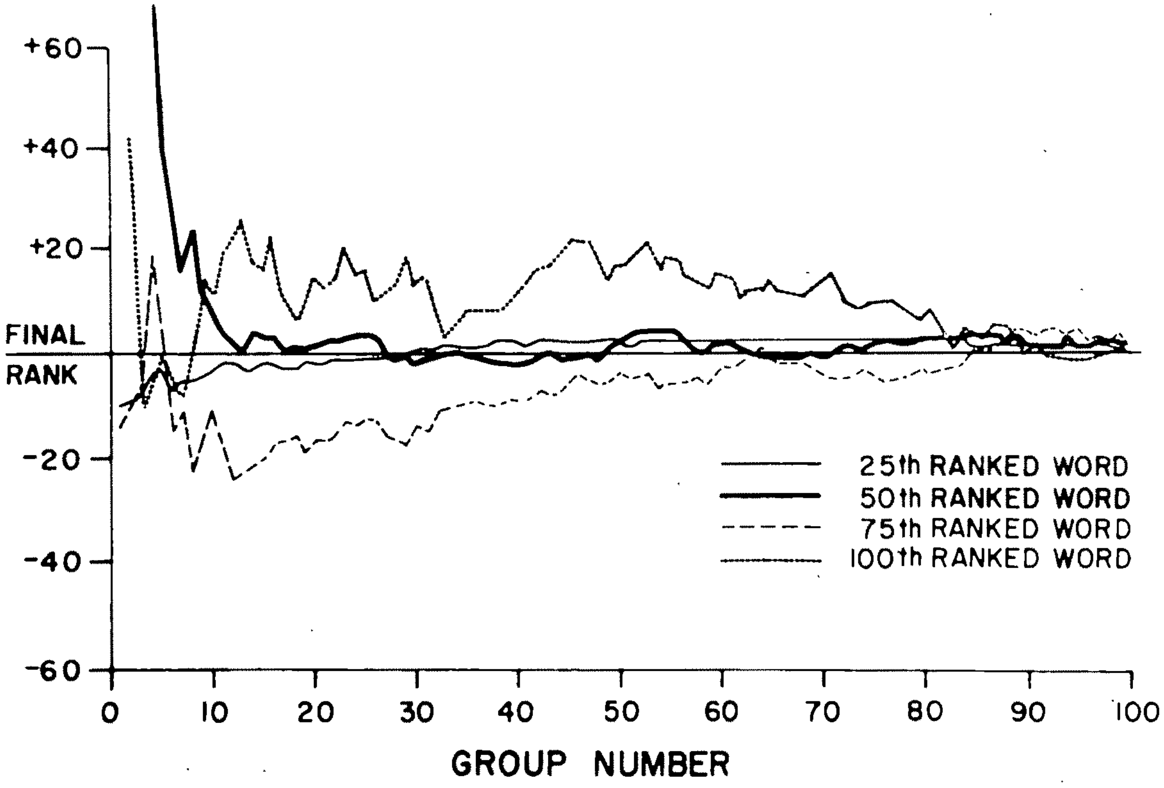


FIG. 3. Rate at which rank of words stabilized

0th ranked word is close to its final rank at the 20,000 ord level, the 75th ranked word takes about 50,000 ords, and the 100th ranked word uses about 80,000 ords to reach a relatively stable position.

Many "key" words are used at such a low frequency hat attempts to improve list searching times by ranking he list by frequency must come after a rather long and edious period of data analysis before their rank can be dequately determined. Thus, sorting lists by individual anks would not result in a significant improvement in omputer efficiency, except possibly the first few very igh frequency words (the first 25 to 250 words). Reatively long lists of words occur with a very low frejuency of use. For example, our analysis found almost ,200 words used only one time in 100,000 words, and over ,500 words used no more than one time per 10,000. These vords account for about 70% of the 3,427 different words ound used.

Although files by rank order appears impractical, the uthors feel there is great value in searching the discard ist by rank *groups*, as is commonly done and as discussed arlier. A rank group, which we like to think of as a ilter level, consists of a group of words which seem to ave nearly the same relative frequency.

6. Summary

The first 100,000 words found in 3,321 reports taken as lictated from radiological (x-ray) patient case records iave been analyzed. These words came from an approximately 5% representative sample of the annual case load n the Radiology Department at the University of rkansas Medical Center.

The rate of introduction of different words as opposed o total words is plotted (broken line) in Fig. 2. Our esults seem to conform to other results reported in the iterature. The rate of introduction of new different vords levels off at about 2,500 words and this occurs when 40,000 to 50,000 words of text have been analyzed. However, unclassified words continue to occur at a significant level of almost 2% (Fig. 2, solid line) as far out as the 100,000 word level. Noise is still at the 1% level which is not tolerable for our application.

Attempts to establish the rank order of words beyond the first several hundred have not met with great success because about 70% of the words appear to occur with such a low relative frequency (no more than one time in 10,000).

Thus, establishing files by rank order appears impractical. On the other hand, filter lists (discard words) by rank groups (words which appear to have near the same frequency) seem quite practical. As currently designed we plan to use a primary filter list of the 25 to 35 highest frequency words and a secondary level filter of 2,500 to 3,500 words. We are now uncertain as to whether it would be best to expand the secondary filter list or to establish a tertiary filter list to reduce the noise level further. We plan to continue the study.

References

1. Barnhard, H. J., and J. M. Long, Computer Autocoding Selecting and Correlating of Radiologic Diagnostic Cases, *American Journal of Roentgenology* (April 1966).
2. Dewey, G., *Relative Frequency of English Speech Sounds,* Harvard University Press, 1923.
3. Booth, A. D., Characterizing Documents—A Trial of An Automatic Method, *Computers and Automation,* November: 32–33 (1966).
4. Schwartz, E. S., Dictionary for Minimum Redundancy Encoding, *Journal of the Association for Computing Machinery,* 10:413–439 (1963).
5. Schwartz, E. S., *Methods of Microglossary Analysis,* 1964 Rochester Conference on Data Acquisition in Biology and Medicine. Publication in progress.
6. American College of Radiology, *Code for Roentgen Diagnosis Indexing,* 1962.

Relationship of Keywords in Titles to References Cited*

Some machine-produced indexes have been compared. The objectives of an index and the search procedure are analyzed. It is proposed that a hypothetical title be used for searching, not a question. This permits comparison of similar items for common characteristics. A relationship between "keywords in title" and "key references" of the same article is given. It is shown that for an efficiently produced article there will be more references than keywords in the title. This is because, according to the basic theorem of linear programming, there will be as many concepts in the article as there are completely utilized references. The structural model proposed requires additional experimental verification.

WM. MANSFIELD ADAMS

Hawaii Institute of Geophysics
University of Hawaii

Comparison of various machine-produced indexes is given in Fig. 1 in terms of the characteristics of the format of the output. The characteristics which are considered, and their definitions, are:

Fixed word length per word: some words are forced to fit into specified word length.

Direct entry: the user enters the main index immediately.

Direct exit: the user passes immediately from the main index to the indexed document (which may be an abstract).

Complete reference: the entire title and location of the document.

Coden: source journal as specified by a few letters.

Screening: minimizing interest in that portion of the title in a KWIC index which appears before the keyword by overprinting with a screen.

Annotation: the addition of words to a title to indicate concepts covered by the paper but not included in the title.

Stop list: list used by the computer as a definition of words that are not to be treated as keywords. Titles will not be entered in the keyword index under these words.

Titles per index page: this ratio is estimated by dividing the total titles treated by the number of pages in the index. The ratio is used only as an indicator and must not be thought of as a value judgment. (Should the erroneous attitude be taken that a high value of this indicator is desirable, then to attain perfection it would only be necessary not to issue an index!) No correction for page size has been made.

Titleless: nowhere in the system is the complete title given.

Detailed discussion of each index is given in Adams (1965).

• General Process of Indexing

The process of searching for information may be diagrammatically illustrated as in Fig. 2. In a very general sense, the user approaches the searching system with certain necessary input information. For the majority of the systems considered in Fig. 1, this is in the form of a list of concepts related to the user's current interests. In the special case of the *Science Citation Index* (2), the input information would be a scientific paper known to be of interest.

In some search systems the user goes directly to the main "scan" index and then to the document indexed in the "scan" subsystem. In other indexes it is necessary that the user compare his input information with a preindex. This serves to translate the input into a form useful in the "scan" index, for example, in the Universal Decimal System. Alternatively, it is possible that the search involves a postindex. That is, the user is directed from the "scan" index to a subsystem having more information concerning the desired document than

* Hawaii Institute of Geophysics Contribution No. 156. This paper is a condensation of a laboratory report (Adams, 1965).

Characteristic of Output / Index	Fixed Word Length	Direct Entry	Direct Exit	Complete Title	Single Journal	Multi Journal	Coden	Screening	Wrap Around	Annotation	Stop List	Titleless	Titles / Index page	Remarks
Rinehart	Y	Y	Y	Y	Y	N	N	N	Y	N	I	N	I	Uses only Sorter
Chemical Titles	N	Y	Y	Y	N	Y	Y	N	Y	N	Y	N		
BASIC	N	Y	Y	N	Y	N	I	Y	Y	Y	N	N		
Assoc. for Computing Machinery (Youden)	N	Y	Y	Y	Y	N	Y	N	Y	N	N	N	10.9	KWIC length of 100+ characters
Meteorological & Geo-astrophysical Titles	N	Y	N	Y	N	Y	N	N	Y	N	N	N	31.6	Post-index
UNIDEK	N	N	Y	Y	N	Y	N	N	I	N	I	N	17.6	Pre-index
Keyword Index to U.S. Government Reports	N	Y	Y	Y	Y	N	N	N	I	N	N	N	10	
Science Citation Index	Y	Y	Y	N	N	Y	N	N	I	I	I	N	U	Legend
MEDLARS	N	Y	Y	Y	N	Y	N	N	I	I	I	N		N = No Y = Yes
Oceanic Coordinate Index	N	Y	N	N	N	Y	N	N	I	Y	I	N		I = Inapplicable U = Unknown
B.S.S.A.	N	Y	Y	Y	Y	N	N	Y	Y	Y	N	N	7.2	

FIG. 1. Comparison of some properties of the output of some machine-produced indexes

is provided by the "scan" system, in particular, the location of the document. A postindex need not be compulsory; if the source location is denoted by the code in the "scan" index, then the user has the option of going directly to the document or going to the postindex. It is quite likely, however, that the user will choose to go to the postindex when the title has been severely truncated in the "scan" index.

Several of the indexes treated here actually index abstracts. Conceptually, the collection of abstracts could be considered as a postindex which gives the user additional information concerning the original article. He may then proceed to the original document or terminate interest, depending on the relevance indicated by the abstract.

• Theory of Information Retrieval

One recent quantification of the information-retrieval process has been given by Goffman (1964). Briefly, the answer to a query is obtained by maximizing an evaluation function of the system output in terms of the probability of relevance of each item of the output

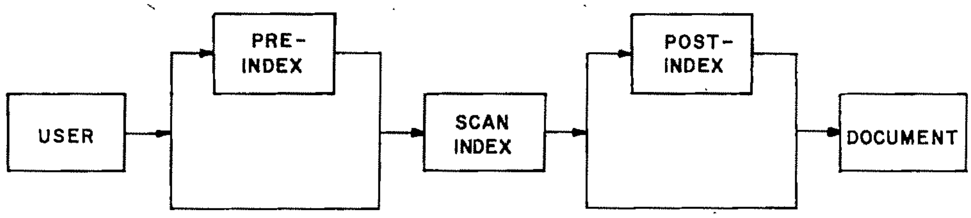


FIG. 2. Diagrammatic flow chart of the searching process

to the query. For this definition of an answer, the necessary and sufficient conditions for a set to be an answer to a question are determined. Although conceptually helpful, this formulation is not operationally oriented because the probability of relevance of every item in the set of documents is not operationally defined. One concept, taken as known, is the critical probability of relevance, which acts as a point of truncation for the sequence of documents as ordered according to decreasing probability of relevance. The function of the information-retrieval system is to locate that subset of documents for which the probability of relevance is greater than the critical probability. Goffman places great weight on the importance of the order in which the documents are presented to the user.

• An Operational Definition of Relevance

We will now develop an operational definition for the term "relevance." (For an alternative and more profound effort to operationally define "relevance," see Hillman, 1964a, 1964b.) We do this by analysis of the searching process. The steps in searching are usually as follows:

1. State the problem in terms of concepts represented by words.
2. Classify the words as significant or insignificant.
3. For some significant word locate all those titles containing that word. This forms a subset.
4. Using each significant word in the problem statement, repeat the above steps and form the collection of subsets.
5. There is now available a subset of titles, say M_1, each one of which contains at least one of the significant words in the problem statement. The number of all such titles will be an integral quantity, say M_1.
6. For some pair of significant words, locate all those titles in M_1 containing both significant words. This forms a subset.
7. Repeat the above step for each pair of significant words.
8. We will now have available a collection of subsets of titles which have at last two of the significant words. We call this the subset M_2. (That this procedure is really used is verified by a recently introduced index.)
9. In a similar fashion, derive the sets M_3, M_4, etc., until the subset is the null set or until the number of included significant words equals the total significant words in tha problem statement.

(Note that the subsets M_1, M_2, M_3, . . . are nested.)

We may then represent the number of titles in each subset versus the number of included significant words as a monotonically decreasing histogram, such as in Fig. 3. We may now interpret a KWIC index in terms of this bar graph and the definition of relevance. Use of the KWIC index is equivalent to starting from the left and progressing to the right. Furthermore, it is necessary to progress *all* the way to the right, that is, the relevance level of every member of the coded set must be determined. A partial ordering by relevance is now evident. Any item in a subset to the right of a given subset is said to be more relevant than any item in the given subset. Consequently, the item having the highest relevance is that having the most number of significant words from the problem statement. Having defined relevance as an objective partial ordering, there no longer seems to be any value of the concept "probability of relevance."

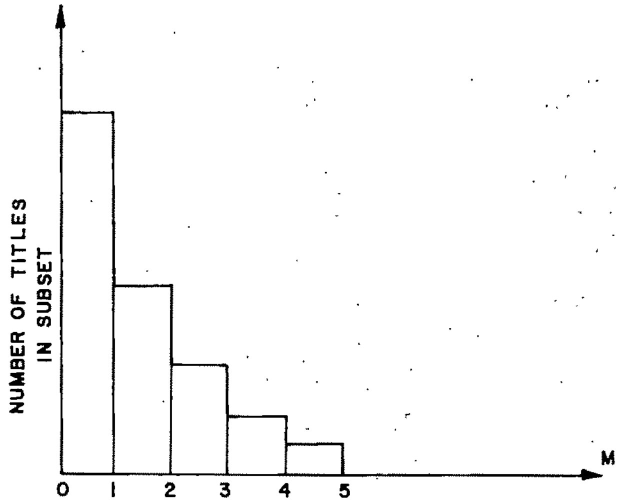


FIG. 3. Schematic histogram illustrating the monotonicity of higher-order subsets. The order of the subset M is defined by the number of keywords which exist in both the model title and the scanned title.

Effects other than relevance need to be included. One we wish to emphasize here is the "assimilation time." Any user interacting with an information-retrieval system will require a definite amount of time to assimilate any one of the documents. This is defined to be the assimilation time. Because a user's time (and memory) are limited, there will be a limit on the number of documents he can assimilate; this might be termed the "assimilation limit." This limitation might truncate a search before the "critical relevance" (corresponding to Goffman's "critical probability of relevance") is reached. We introduce a symbology similar to that of Goffman. Following this notation, with appropriate modifications and extensions,

$S =$ set of source documents
$q =$ question
$u =$ user
$x =$ element of coded set for S
$r_q =$ relevance of S(x) to q (the relevance of x to q will be used as an estimate of r_q)
$R_u =$ critical relevance
$t_u =$ time required by the user to assimilate S(x)
$T_u =$ assimilation limit of the user.

Introduction of the time element requires elaboration f Fig. 3. Assuming the distribution of assimilation ime to be approximated by a Poisson distribution, we ive such an elaboration in Fig. 4 for a continuous nd finite distribution with respect to both relevance nd assimilation time. Now each document S is characerized by a code, x, a relevance dependent on the uestion and estimated by r_q, and the assimilation time $_u$, dependent on the user. The objective will be to maximize the number of most relevant documents assimilated ithin the assimilation limit, or the critical relevance, .e., maximize:

$$Z = \sum_{i=1}^{m} S_i(x, r_q, t_u)$$

ubject to: $r_q > R_u$

nd $\sum_{i=1}^{m} t_u < T_u$

This means, graphically, starting at the "tail" and ssimilating into the "hill," and, for documents of equal elevance, always using documents of short assimilation ime before documents of long assimilation time. Note wo features. First, the searching procedure might be mproved if the user approached the retrieval system ith a hypothetical title instead of a question. This s suggested because the code for each of the documents eing searched is a title. The hypothetical title should e so posed that the user would expect a paper having hat title to provide him with the answer to his probem. This type of retrieval process shifts additional ork to the user, but the increased efficiency in using he index might be adequate justification. Verification of this hypothesis is provided by semantically analyzing the procedure of Lancaster and Mills (1964). Although they repeatedly describe the searching process as using "questions" for input, they give as an example (p. 12): "We now search these [uniterms] one at a time; e.g., a question on 'flow solutions for chemically reacting gas mixtures' would be searched for first. . . ." Note that the "question" is presented as a "hypothetical title." Nor is this an exception (see p. 6).

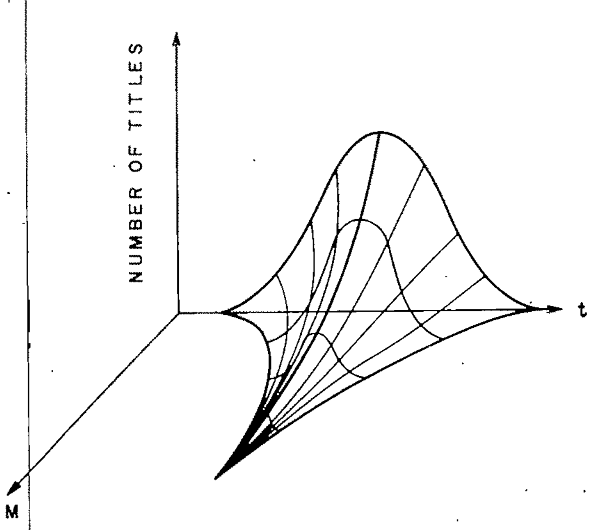


FIG 4. Graphical illustration of general relationship among he assimilation time, t, the order of the subset, M, and he number of titles. The "tail" consists of those titles havng the highest value of M; the "head," of those having the owest value of M.

Secondly, it should also be noted that the operational definition given here for relevance can be refined. That is, instead of defining relevance by number of included significant words in the title, the user forms a hypothetical abstract. The abstracts are used as the codes to the set of documents. The hypothetical abstract is compared with each abstract. As an extremism, this approach can be expanded to include the entire article. Such does not appear to be feasible with the current computing equipment (nor with the usual reluctance of users to precisely formulate the input to a retrieval system).

• Relationship of Title Words and References

Of the several types of indexes reviewed in Fig. 1, the *Science Citation Index* appears to differ most from the others. The differences among the various types of indexes are important because index makers sometimes make more than one type of index for a given set of documents. Because the assimilation limit is essentially a time limit, the users must then make a decision concerning which index, or how many indexes, will be used to search the existing documents. In making this decision, it is worthwhile to realize the relationships between the different types of indexes. Here we will discuss the relationship of the title-oriented index to the reference-oriented index. We utilize the theory of linear programming.

We consider the author of an article during the entire process of creating the article. Assume that the author will operate at optimum efficiency. One of the resources of the author is time, T. This time may be expended in a variety of activities, each of which contributes to the resulting article. Let us designate the concepts included in the resulting article as C_1, C_2, . . . C_n. For mathematical convenience, we quantize "concept" and consider it to be composed of "conceptlets." The activities that the author may perform in furtherance of the creation of the paper are several: we designate these as X_1, X_2, X_3, . . . X_m. For example, the first activity might be reading document "one"; the second activity, reading document "two"; etc.

Now each activity, performed for a unit time, will contribute a various number of conceptlets to each of the several concepts. Thus, from performing activity one for a unit time, the author will obtain a_{11} "coneptlets" related to C_1, a_{12}, related to C_2, etc. Let the mount of time spent on X_j be the intensity x_j. Then, 1 general,

$$C_i = \sum_{j=1}^{m} a_{ij}x_j (i=1, 2, \ldots n)$$

'here will be a limit as to the time which can be pent in any particular activity and the productivity emain directly proportional to the time. We limit this ime to be less than y_j. By assumption, only linear reponse is considered.

Furthermore, all the time spent on all the activities ıust be less than, or equal to, the time available, T.

The author will consider the various concepts to have ertain relative values, say u_1, u_2, etc. (For numerical alues, we might just use the number of words he would ke to write on that concept.)

Based on the foregoing ideas and terminology, we ıay formulate the following linear-programming probım. Maximize the objective function

$$Z = u_1C_1 + u_2C_2 + \cdots + u_nC_n$$

Vhere:

$$C_i = \sum_{j=1}^{m} a_{ij}x_j \; (i=1, 2, \ldots n)$$

ubject to conditions:

$$\sum_{j=1}^{m} x_j < T$$

$$x_j < y_j (j=1, 2, \ldots m)$$

Jnder known conditions this problem will have a olution. The important aspects here are not necesarily the detailed solution, but the general properties f the solution. We consider these now.

Applying the basic theorem of linear programming, re find that *under an optimum arrangement, there will e as many concepts in the created article as there are eferences completely utilized* (see Baumol, 1963). Thus, ' the author desires to include n concepts, only n eferences will be used to capacity, in general. In ther words, in general, *in an efficiently produced article, here will be more references than concepts.*

We have found a relationship between the concepts overed by an article and the references on which the rticle is based. Our intention is to derive a relation etween title words and the references; therefore we eed a relationship between title words and concepts. 'his is provided, in an empirical sense, by the *Oceanic 'oordinate Index,* under its "Citations" index. In ddition to the title, annotation is given. We estimate hat there are approximately as many concepts as there re keywords in the title. Each keyword represents one oncept, by assumption.

Using the foregoing findings, we can relate the keyords in the title of an article to the references of that article. We predict that *for a given article, efficiently produced, there will be more references than keywords.* This expectation has been checked using estimates from articles in scientific journals. The actual ratio is probably about three references to one keyword. Inspection reveals that the references are of two types, those pertaining to data, and those pertaining to concept.[2] We estimate that about half the references pertain to data, hence, to a first approximation, we estimate that:

$$\frac{\text{number of conceptual references}}{\text{number of keywords-in-title}} = \frac{3}{2}$$

Defining a "key reference" to be a "conceptual reference" which is fully utilized, we may hypothesize:

$$\frac{\text{number of key references}}{\text{number of keywords-in-title}} = 1.$$

Additional investigation is necessary to confirm the foregoing estimates and to develop operational definitions of the terms.

Note that the foregoing treatment has implied that prior documents (the references) are the only source of material for constructing a concept. This limitation has facilitated the presentation. Actually, of course, "nature" may be treated as a "document" and new data (facts) or time-space relationships attributed to that reference. "Nature" might appear in the bibliography in the form of a laboratory notebook, field book, etc.

Note that the situation may be nonlinear. Consideration of the extreme cases makes this possibility immediately apparent. Thus, one would intuitively expect a book or review article with a very short title, such as "Technical Libraries," to have more references than one with more significant words in the title, such as "Indexing of Abstracting Services in Technical Libraries." At the other extreme, a paper dealing with such a restricted subject that the only reference is to one of the author's own previous papers might be expected to have an extremely long title. In order to properly circumscribe the very detailed topic being covered, we might hypothesize that the relationship between the number of references and the number of keywords is as shown in Fig. 5. Indeed, this relationship could be used to define the nature of a document. For example, a document falling between the origin and K_1 would be a *book,* between K_1 and K_2 would be an *article,* and a document lying beyond K_3 might be defined as an *exercise.* Items falling in the K_2–K_3 range would be *article-exercises.* This illustrative notation is not suggested for actual application without further study to confirm our hypothesis about the shape of the basic relationship between keywords and key references.

Another theorem of linear programming is especially

[2] A similar division is made by Lancaster and Mills (6, p. 9) in discussing the applications of a law relating "recall" and "relevance."

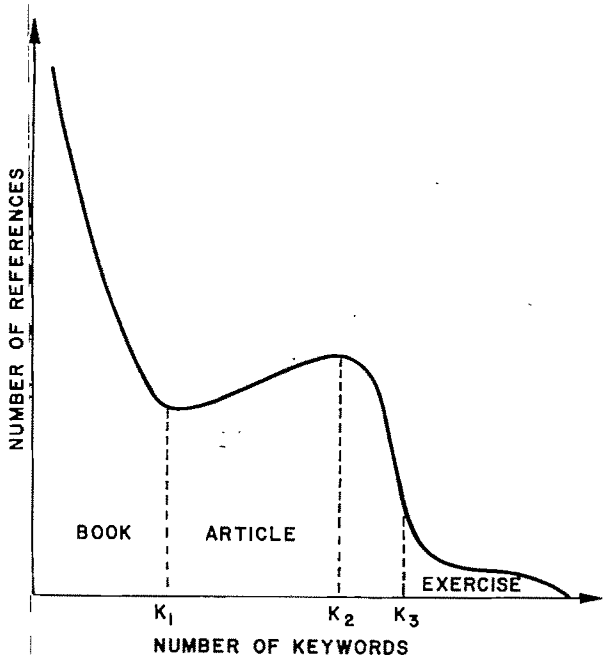


FIG. 5. Hypothetical relationship of references to keywords for books, articles, or exercises

applicable to solving our earlier dilemma over which of the criteria—relevance or assimilation time—would be dominant. It is shown in linear programming theory that "a program is the most efficient if and only if it contains included activities such that no excluded activity contributes more to the objective function than an equivalent combination of included activities" (Dorfman, Samuelson, and Solow, 1958, p. 164). Thus, the interrelationship of the assimilation time (the intensity, x_j) and the relevance (the productivity, a_{ij}) becomes clearer by assuming the existence of relative value (u_i). An expression for the marginal value of the created article in terms of the source references may be obtained from Farka's theorem (Dorfman, Samuelson, and Solow, 1958, p. 191). However, there seems little point in giving the details here since the relative values are defined subjectively.

The expansion of the foregoing interpretation to nonlinear and dynamic conditions is relatively straightforward.

This work reports on the format. As evidenced by *Index Medicus,* the format is highly machine-dependent. For information about development of instruments, machines, and equipment for data storage and retrieval, such as are relevant to indexing, see Elliott (1964).

• Conclusions

An operational definition for relevance has been given in terms of the concepts in the posed question included in the title of the document.

The searching process is outlined to be:

1. State the problem in terms of concepts represented by words.
2. State hypothetical title of a hypothetical article which would be expected to answer the posed question.
3. Classify the words as significant or insignificant.
4. For some significant word, locate all those titles containing that word. This forms a subset.
5. Using each significant word in the problem statement, repeat the above steps and form the collection of subsets.
6. There is now available a subset of titles, say M_1, each one of which contains *at least* one of the significant words in the problem statement. The number of all such titles will be an integral quantity, say M_1.
7. For some pair of significant words, locate all those titles in M_1 containing both significant words.
8. Repeat the above step for each pair of significant words.
9. There is now available a subset of titles which have *at least* two of the significant words. This we call subset M_2.
10. In a similar fashion, derive the sets M_3, M_4, etc., until the subset is the null set or until the number of included significant words equals the total significant words in the problem statement.

Use of a hypothetical title for searching a list of titles permits comparison of similar items for common characteristics.

A proof is cited that relates the references to the concepts covered by an article. The relation of key references to keywords is then empirically estimated by ascertaining the keyword-concept relationship.

Goffman has noted the possible significance in the order in which the documents are presented to the user. Probably much more important is the relative order of importance of keywords in a title. If authors were to order the keywords by order of importance rather than by grammatical rules, the hypothetical title could be similarly ordered. Comparison could then be made for the order of the joint keywords.

• Acknowledgments

Assistance in this study phase of the work, as opposed to the production phase, has been provided by Miss Joyce Wright, Miss Virginia Crozier, Miss Genevieve Correa, and Mr. James Fitzpatrick, of Gregg Sinclair Library, University of Hawaii. Discussion with Professor John Wise, of the Department of Economics, University of Hawaii, concerning the application of linear programming to the creation of a scientific article has been most helpful. However, the author is responsible for the opinions set forth here.

Financial support has been provided by the National Science Foundation under NSF grants N-1272, GN-95, and GP-3473.

This article is a condensation (essentially an excerpt) of an institute report (Adams, 1965).

Bibliography

Adams, W. M. *A Comparison of Some Machine-Produced Indexes,* Hawaii Institute of Geophysics Report 65–1, 1965.

B.A.S.I.C. *Biological Abstracts Information Dissemination System,* Biological Abstracts, Inc., Philadelphia, Pa. (monthly), April 1, 1962–date.

Baumol, W. J. *Economic Theory and Operations Analysis,* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963.

Chemical Titles, The American Chemical Society, Easton, Pa. (biweekly), 1960–date.

Dorfman, R., P. A. Samuelson, and Robert M. Solow, *Linear Programming and Economic Analysis,* McGraw-Hill Book Co., Inc., 1958.

Elliott, J. Richard, Jr. Information Please, *Barron's,* I: p. 3 (Oct. 19, 1964); II: p. 3 (Nov. 2, 1964).

Geoscience Abstracts, special supplement: Unidek, Kwic, and Arthur Indexes. American Geological Institute, Washington, D. C., 1964.

Goffman, W. A Searching Procedure for Information Retrieval, *Information Storage and Retrieval,* 2:73–78 (1964).

Hillman, D. J. The Notion of Relevance (1), *American Documentation,* 15 (No. 1):26–34 (1964a).

———. The Notion of Relevance (2), *American Documentation,* 15 (No. 2) (1964b).

Index Medicus, National Library of Medicine, U. S. Superintendent of Documents, Washington, D. C. (monthly), 1960–date.

Keyword-In-Context (KWIC) Indexing, *General Information Manual E20-8091,* International Business Machines Corporation, White Plains, New York, 1962.

Keyword Index to U. S. Government Technical Reports, U. S. Dept. of Commerce Office of Technical Services, Washington, D.C. (biweekly), Vol. 1 (June 15, 1962–date).

Lancaster, F. W., and J. Mills. Testing Indexes and Index Language Devices: The ASLIB Cranfield Project, *American Documentation,* 15 (No. 1):4–13 (1964).

Meteorological and Geoastrophysical Titles, American Meteorological Society, Boston, Massachusetts (monthly), 1950–date.

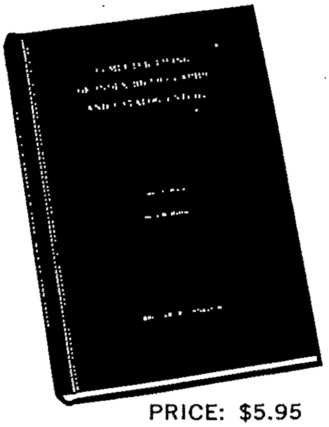
Oceanic Coordinate Index, Mission Bay Research Foundation of San Diego, La Jolla, California (bimonthly), 1964–date.

Rinehart, W. A. Personal correspondence, 1962.

Science Citation Index, Institute for Scientific Information, Inc., Philadelphia, Pa. (quarterly and annual), 1961, 1964–date.

Seismological Society of America. Bulletin, San Francisco, California (bimonthly), 1911–date.

Youden, W. W. Indexing to Journal of the Association for Computing Machinery, *J. Assn. Computing Machinery,* 1–10 (1963–1964).



A Multiple Testing of the Natural Language Storage and Retrieval ABC Method: Preliminary Analysis of Test Results[*]

After a brief summary of the test program, the statistical results tabulated as over-all "ABC-Relevance Ratios" and "ABC-Recall Figures" are presented and reviewed. An abstract model developed in accordance with Max Weber's "idealtypus" is used in discussing such observations as the absence of the detrimental effects of an inverse relationship of Relevance and Recall upon a system's effectiveness. The increase of Recall in proportion to the number of documents retrieved is attributed to the ABC-system's peculiar capability of making the user an integral part of the system.

BERTHOLD ALTMANN

Harry Diamond Laboratories
Washington, D. C.

Introduction

The principles and operations of the ABC storage and retrieval method were briefly outlined in a presentation at the 151st National Meeting of the American Chemical Society (1),[1] and the preparations and the procedures of the test performed to assess the capability of the method were the subject of a detailed technical report (2).[2] In this report we discussed the test program and its principles, the selection and organization of the collection, the preparation and standardization of the requests, the procedures and forms used in the retrieval operations to record results, the methods of evaluating the data, and the transfer of the evaluated data to the summary sheets.

In the present article we will deal primarily with a preliminary analysis of test results.[3] A mathematical model and the final statistical analysis of the accumulated data are being processed for publication. However, the information presented in this factual report is consistent with the conclusions of reference (3).

Table 1 briefly outlines the organization of the test operations. Two types of requests were prepared by 41 HDL scientists and engineers (Group 1 of Group I): (a) 225 requests based on the contents of papers randomly selected from the entire test collection; and (b) 36 requests based only upon a general knowledge of the subject areas covered by the collection. The control group (Group II) consisting of 31 senior scientists and engineers (including members of DOD, Air Force and Navy research agencies, and the National Bureau of Standards) standardized all requests with respect to form and content, reducing the number of document-based requests from 225 to 100 in the process.

For the retrieval, Group 1 was divided into two teams (1a and 1b). Each team processed 50 requests that its own members had helped to formulate and 50 requests formulated by members of the other team. This procedure was used to determine if any bias was introduced by having an operator retrieve information in response to requests he himself had formulated. In addition, both teams processed the 36 freely styled requests. Also used

[*] I am indebted to Dr. Maurice Apstein, Messrs. Theodore B. Godfrey and Hoyt W. Sisco in HDL, and to Mr. Martin H. Weik, Jr., in ARO for valuable suggestions, and to Mr. Hoover Ogata for the critical comments that prompted the inclusion of additional evidence and a stronger formulation of the presented thesis. Discussions with Mr. F. W. Lancaster and an exchange of written communications with Mr. Cyril Cleverdon have contributed to a better mutual understanding; they crystallized differences of opinions, in particular those based upon the basic variations of the systems tested and the HDL plan of evaluating the variables of the first-generation ABC model for the express purpose of gaining the information necessary to design a system that can satisfy the user. Finally, I acknowledge with gratitude the help and advice given to me by Mr. R. A. Fairthorne.

[1] The presentation is already dated in several respects. For example, the plan of introducing subject schemes to organize large clusters of descriptive phrases had to be abandoned for a more practical and economic solution.

[2] The Army Research Office, Scientific and Technical Information Division, Washington, D. C., supports the development of the ABC-System as a contribution to the Army Technical Library Improvement Studies.

[3] Organization and interpretation of the tabulated data were subject to a continuing review following standard statistical procedures, a task diligently and resourcefully handled by W. H. Menden, who, in turn could rely on the advice and guidance of B. M. Kurkjian.

TABLE 1. Test Outline

	Group I				Group II *
	Group 1 †		Group 2 ‡	Group 3 §	
	1a	1b			
I. Preparation of Requests:					
Type a: Document-based	225				
Type b: Freely styled	36				
II. Standardization of Requests:					
Type a reduced to:					100
Type b:					36
III. Test Runs:					
Type a: Own requests	50	50			
Type a: Counterpart's requests	50	50	100	100	
Type b: Requests	36	36	36	36	
Total	136	136	136	136	
IV. Pre-evaluation of results	136	136	136		
V. Final evaluation of results					136

* 31 Senior scientists and engineers including those of other agencies
† 41 HDL scientists and engineers
‡ 6 Analysts (George Washington University)
§ 6 HDL Librarians

in retrieving were Groups 2 and 3 of Group I. Group 2 included six professors from George Washington University, who took the place of HDL branch and laboratory chiefs, since the latter were not available for the experiment. Group 3 included six members of the HDL Library Staff.

While the first group (two teams) of 41 bench workers represented 77 percent of the population of operators and each of its members was assigned about 5 percent of the total test requests according to their individual interests, the remaining two groups (2 and 3) each represented 11.5 percent of all operators, and individual members were assigned 16 to 17 per cent of all test requests. This distribution of operators and requests was intended to simulate a realistic situation, where a bench worker studied and probed a relatively narrow subject area, while a supervisor or librarian covered a multiplicity of subjects.

The operators used three tools for retrieval: (1) the ABC method with a short dictionary (Version I); (2) the ABC method with a long dictionary (Version II); and (3) the KWIC title list prepared for all documents included in the collection. The ABC dictionaries were KWIC-type listings of index phrases describing the contents of the collection and differed merely in the amount of detail displayed. The KWIC title list was used for comparison as well as a control.

An operator processed the same request in successive runs with three different tools with individual test runs normally separated by a lapse of a day. All operator groups tested the short ABC dictionary (Version I) first; normally, a retrieval with the KWIC title list followed, and Version II was tested last. To determine any operator bias, Group 1b tested ABC Versions I and II in order, using the KWIC title list last.

For the determination of r, the number of relevant items in the collection for each request, the members of Group II followed a fixed procedure designed to guarantee a high degree of objectivity. Essentially, this process involved the following three steps:

1. Analysis of the wording of the requests to form linguistic building blocks (concepts) and asserting the minimum combinations of these that would be acceptable.
2. Determination of the total number of different documents retrieved for each request that were relevant by the above criterion.
3. Determination, if possible, of additional relevant items not retrieved by the test operators.

To determine the relevancy of an item retrieved in response to a particular request, the evaluators established a ranked list of linguistic building-blocks for the request; each block represented a conceptual unit. These conceptual units were then compared with the contents of the document, and relevancy was assigned if, of several predetermined combinations of conceptual units, one was found to be descriptive of the document.

In the next step, the total number of *different* relevant items retrieved in response to that request was determined by comparing the 12 sets of relevant documents retrieved for each request (all requests were tested by *four* operators using three different tools, the two ABC Dictionaries Versions I and II, and the KWIC list).

In addition to this, the members of Group II used various lengthy and sophisticated retrieval tools and procedures not available to the test operators to determine if relevant items were in the collection that had not been retrieved. In particular, they checked a systematic card catalog[4] consisting of abstract cards with cross references and supplemented by an alphabetical index. Members of Group II also employed specially constructed retrieval loops (2, pp. 11, 16–17) as control mechanisms to check for further relevant materials. Finally, the quantity r was obtained as the sum of the different relevant items found by the test operators and the members of the control Group II.

With a value of r asserted for each request and the quantities x (number of relevant items retrieved) and n (number of items retrieved) determined for each retrieval run, individual Relevance Ratios (sometimes called Precision Ratios) defined as x/n and Recall Ratios defined as x/r were calculated. For statistical reasons, averages of these ratios (for groups of observations) were calculated by averaging the numerators and denominators separately, i.e., for k observations of $x_i/n_i (i=1 \,..\, k)$, we obtained:

$$(x/n)_{av} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} n_i}$$

Among the averages obtained were those for all 136 requests processed by a given group of retrieval operators, or for all four groups of operators using either version of the ABC dictionary (see Tables 3 and 4). Statistical tests (χ^2 analysis were made to insure the validity of these averages; the results are recorded in Table 2. The tests produced no evidence that average Relevance Ratios obtained for different sets of data could not be validly combined.

Of the average Recall Ratios, however, only those representing the total (Versions I and II) results for a given group of testers and a given set of questions were found suited to form meaningful combinations. To stress the fact that average Recall Ratios as well as their combinations are of the limited significance, we will refer to them in the subsequent discussions as average Recall Figures.[5]

Furthermore, based upon evidence of the test (Table 2), we assumed the existence of an ABC system parameter for Relevance. The parameter which is called "ABC-Relevance" in this article will be estimated by the grand total average over all Relevance Ratios observed for the ABC system.

For the same reason, we cannot assume the existence of a similar systems parameter for Recall, to be estimated by averaging Recall Ratios. Although we have consistently calculated averages and have called them Recall Figures, we have done so only to facilitate a discussion of the problems involved.

TABLE 2. Results of Statistical Tests Performed to Justify the Combination of the Different Sets of Data

(All decisions were made at a level of significance of 5%)

I. Is the combination of the four sets of data obtained for each retrieval tool from the four groups of retrieval operators acceptable?

Method: X^2 analysis on a four by two contingency table.

	ABC Dictionary			
	Version I		Version II	
	Relevance	Recall	Relevance	Recall
36 Fully styled requests	yes	no	yes	no
100 Source document requests	yes	no	yes	no

II. Is the combination of the two sets of data obtained from each group for each set of requests in testing ABC Versions I and II acceptable?

Method: Computation of the standard error of the average difference.

	Relevance Ratio	Recall Ratio
36 Fully styled requests	yes	yes
100 Source document requests	yes	yes

III. Is the combination of the two sets of corresponding data obtained from the 36 freely styled and the 100 document-based requests acceptable?

Method: Computation of the standard error of the differences.

	Relevance Ratio	Recall Ratio
Short Dict. (ABC Version I)	yes	no
Long Dict. (ABC Version II)	yes	no
Combination of I and II	yes	no

The employment of four different operator groups, three retrieval tools, two retrieval sequences (I, KWIC, II; I, II, KWIC), and essentially three types of requests (freely styled, document-based "own" and document-based "other") produced a complex test result with four variables specifying each run. If the distinction between document-based "own" and "other" is disregarded, since they compensate for each other in a comparison of Groups 1a and 1b, there still remain four groups (1a and 1b distinguished by the retrieval sequence), three tools, and two types of requests, which result in 24 sets of data.

Before we enter into the discussion and interpretation of the results we must raise the question of the general usefulness of validity of test data obtained for a particular system in a particular test environment. Without a clear understanding of (1) the systems themselves, (2) the testing methods (by necessity adjusted to evaluate the operational capability of the individual systems),

[4] To obtain this catalog, the test collection had to be indexed under the classification of "Solid State Abstracts."

[5] The problem of arriving at a Recall Ratio that is representative of the test or the ABC method is analyzed in reference (3).

Table 3. Average Relevance Ratios and Recall Figures for 36 Freely Styled Requests

	ABC Version I (Short)				ABC Version II (Long)				I and II Combined				KWIC Title List			
	Relevance		Recall		Relevance		Recall		Relevance		Recall		Relevance		Recall	
	x/n	%	x/r	%	x/n	%	x/r	%	x/n	%	x/r	%	x/n	%	x/r	%
Group 1a I→KWIC→II	$\frac{52}{57}$	91.2	$\frac{52}{269}$	19.3	$\frac{53}{58}$	91.4	$\frac{54}{275}$	19.6	$\frac{105}{115}$	91.3	$\frac{106}{544}$	9.15	$\frac{62}{97}$	63.9	$\frac{62}{278}$	22.3
Group 1b I→II→KWIC	$\frac{55}{60}$	91.7	$\frac{55}{261}$	21.1	$\frac{69}{80}$	86.3	$\frac{69}{278}$	24.8	$\frac{124}{140}$	88.6	$\frac{124}{539}$	23.0	$\frac{59}{83}$	71.1	$\frac{59}{278}$	21.2
Group 2 I→KWIC→II	$\frac{71}{83}$	85.5	$\frac{71}{278}$	25.5	$\frac{82}{90}$	91.1	$\frac{82}{278}$	29.5	$\frac{153}{173}$	88.3	$\frac{153}{556}$	27.5	$\frac{65}{93}$	69.9	$\frac{65}{278}$	23.4
Group 3 I→KWIC→II	$\frac{75}{87}$	86.2	$\frac{74}{259}$	28.6	$\frac{71}{82}$	86.6	$\frac{72}{246}$	29.3	$\frac{146}{169}$	86.4	$\frac{146}{505}$	28.9	$\frac{46}{72}$	63.9	$\frac{46}{246}$	18.7
Group average*	$\frac{253}{287}$	88.2	$\left(\frac{63}{267}\right)$	(23.6)	$\frac{275}{310}$	88.7	$\left(\frac{69}{269}\right)$	(25.7)	$\frac{528}{597}$	88.4	$\left(\frac{132}{536}\right)$	(24.4)	$\frac{232}{345}$	67.3	$\frac{58}{270}$	(21.5)
Team average†		[84.5]	$\left(\frac{154}{278}\right)$	[55.4]		[87.8]		[59.7]		[86.2]		[78.4]				

*Figures in parenthesis are combinations shown to be not valid. (See section titled "ABC Relevance and Recall Figures.")
†Team average is the accumulated result of 1a, 1b, 2, and 3, with all duplications eliminated and with each discrete document counted in the accumulated x.
Note: Although r is theoretically constant (278), it is reduced in several sets of data because some runs were disqualified for technical reasons.

Table 4. Average Relevance Ratios and Recall Figures for Document-based Requests

	ABC Version I (Short)				ABC Version II (Long)				I and II Combined				KWIC Title List			
	Relevance		Recall		Relevance		Recall		Relevance		Recall		Relevance		Recall	
	x/n	%	x/r	%	x/n	%	x/r	%	x/n	%	x/r	%	x/n	%	x/r	%
Group 1a I→KWIC→II	$\frac{192}{221}$	86.9	$\frac{192}{846}$	22.7 *[46.5]	$\frac{193}{220}$	87.7	$\frac{193}{847}$	22.8 [48.5]	$\frac{385}{441}$	87.3	$\frac{385}{1693}$	22.7	$\frac{212}{276}$	76.8	$\frac{212}{860}$	24.7 (58)
Group 1b I→II→KWIC	$\frac{134}{161}$	83.2	$\frac{134}{860}$	15.6 [41]	$\frac{150}{166}$	90.3	$\frac{150}{857}$	17.5 [50.5]	$\frac{284}{327}$	86.9	$\frac{284}{1717}$	16.5	$\frac{158}{232}$	68.1	$\frac{158}{860}$	18.4 [46]
Group 2 I→KWIC→II	$\frac{202}{241}$	83.8	$\frac{202}{860}$	23.5 [50]	$\frac{196}{232}$	84.5	$\frac{196}{860}$	22.8 [54]	$\frac{398}{473}$	84.1	$\frac{398}{1720}$	23.1	$\frac{201}{267}$	75.3	$\frac{201}{848}$	23.7 [ ]
Group 3 I→KWIC→II	$\frac{110}{126}$	87.3	$\frac{110}{844}$	13.0 [35.4]	$\frac{168}{202}$	83.2	$\frac{168}{855}$	19.6 [48]	$\frac{278}{328}$	84.8	$\frac{278}{1699}$	16.4	$\frac{137}{188}$	72.9	$\frac{137}{836}$	16.4 [ ]
Group Average†	$\frac{638}{749}$	85.1	$\left(\frac{638}{3410}\right)$	(18.7) ([43.3])	$\frac{707}{820}$	86.2	$\left(\frac{707}{3419}\right)$	(20.7) ([50.3])	$\frac{1345}{1569}$	85.7	$\left(\frac{1345}{6829}\right)$	(19.7)	$\frac{708}{963}$	73.5	$\left(\frac{708}{3404}\right)$	(20.8) ([ ])

*Figures in brackets are recall percentages based on the retrieval of the source document only.
†Figures in parenthesis are combinations shown to be not valid. (See section titled "ABC Relevance and Recall Figures.")
Note: Although r is theoretically constant (860), it is reduced in several sets of data because some runs were disqualified for technical reasons.

(3) the precautions taken to eliminate bias, and (4) the processes of evaluating different elements, in particular the satisfaction of the different user groups, a comparison of the systems and their performance will be futile despite such common designations as Relevance and Recall given to the measuring rods.

In order to preclude possible misuse and misinterpretation of our statistical results, I repeat here briefly the description of the ABC system's characteristic features and of the methods employed in the evaluation process although the detailed explanations had been given in our preceding report (2, pp. 1–2, 22–25).

The system provides the scientist with a display of organized, self-explanatory descriptive phrases as well as a comprehensive basis for detecting associations between subjects, disciplines, and ideas pertaining to his problem. Anticipating a future information system where a scientist will directly confer with the text (stored in a computer) by means of communication links, remote teletype consoles, disc memories, and optical displays, the ABC dictionary provides the means for direct communication with the contents of a collection. The retrieval operations, the selection of keywords in accordance with an initial formulation of the request, the matching of words and phrases, the follow-up of leads to additional clusters of descriptive phrases (as the searcher evolves a better strategy) are admittedly complex processes, but they are as rapid as a knowledgeable investigator can make decisions while scanning the appropriate pages of the ABC dictionary. Whatever system he is using, the responsibility for making final decisions is inevitably the investigator's. However, in the ABC system, most of these decisions are made during the initial phases of the retrieval process.

Because the ABC system provides the scientist with complete freedom to select, reject, and browse, we would have distorted the operational realism had we imposed upon our operators an upper or lower limit for the numbers of documents to be withdrawn from the test collection or for the length of time to be spent in a retrieval run. Our primary objective was not a performance assessment of individuals or groups, but an evaluation of the system, its performance and its capability. Since the intensity of an operator's effort is bound to affect an individual result, the relative number of items recalled may be an appropriate measure of this human factor; and with its effect on test results known, it is possible to discuss the capability of the system.

• Test Results

The average Relevance Ratios and Recall Figures for the 24 sets of data and for their combinations were calculated. They are tabulated in Tables 3 and 4 for the freely styled and document-based requests, respectively.

The average Relevance Ratios for Version I of the ABC dictionary range between 85.5 percent and 91.7 percent (88.2 percent total average) for the freely styled requests; and between 83.2 percent and 87.3 percent (85.1 percent total average) for the document-based requests. The corresponding results for Version II are 86.3 to 91.4 percent (88.7 percent average) for the freely styled requests; and 83.1 to 90.3 percent (86.2 percent average) for the document-based requests. The average of both versions obtained with freely styled requests is 88.4 percent, and with document-based requests, 85.7 percent. The total average using all data which we consider an estimate for the ABC Relevance, a system parameter, amounts to 87.1 percent.

A brief glance at the average Recall Figures shows proportionately wider ranges from 19.3 to 28.6 percent, from 19.6 to 29.5 percent, from 13.0 to 23.5 percent, and from 17.5 to 22.8 percent. If we were permitted to average these, we would obtain 23.6, 25.7, 18.7, and 20.7 percent, or 22.8 percent for all data.

A mathematical model (based on the probability distribution functions of the variables), with confidence limits for the final ABC Relevance Ratio and ABC Recall Figures are discussed in detail in reference (3).[6]

• ABC Relevance and Recall Figures

The ABC Relevance as estimated by the average of all Relevance data (87.1 percent) is high and the ABC Recall Figures, arrived from all Recall data (22.8 percent), is low. This might be interpreted to indicate an inverse relationship of one to the other. However, the analysis of our results shows that such a conclusion is not appropriate. As a first indication, when we (following the method of the first Cranfield Test) calculated the average Recall Figure for the returns from the set of 100 requests on the basis of the retrieval of the source document, we obtained the following scores (Table 4): 46.5, 41, 50, and 35.4 (43.3 average) percent for Version I; 48.5, 50.5, 54, and 48 (averaging 50.3) percent for Version II; and an over-all average for the two versions of 46.8 percent. Moreover, comparison of average Relevance Ratios and Recall Figures for Versions I and II and the KWIC title list simply do not bear this hypothesis out. The decrease in average Relevance Ratio with the KWIC system versus either ABC version is accompanied by no significant increase in average Recall Figure.

TEAM RESULTS

Similar trends were found in the analysis of the team results (Table 3) for the freely styled requests. Because of the great expenditure of time and money, the analysis of team effort was limited to the freely styled requests, and we can propose no reason why the results should be different for the set of 100 requests. In calculating the team results, all distinct items retrieved by the four opera-

[6] Confidence bands for Relevance and Recall were determined at a level of 95 percent. Recall Ratios, however, could be established only for the data of the constant r subsets.

tors concerning a given request were summed to obtain n, the number of documents retrieved; all duplications were not counted. The number of documents that were found relevant were counted. These revised figures were then used to calculate team Relevance Ratios and Recall Figures which are given in brackets in Table 3.

The average Recall Figures obtained in this manner rose from 23.6 to 55.4 percent for Version I, from 25.7 to 59.7 percent for Version II, and from 24.7 to 78.4 percent for Versions I and II combined (for a team of four making two measurements each). Average Relevance Ratios calculated similarly decreased only to 84.5 percent from 88.2 percent, to 87.8 percent from 88.7 percent, and to 86.2 percent from 88.4 percent. The first improvement in Recall Figures, either for Versions I or II, requires that the individual operators produce quite different but equally pertinent responses to the same request using the same system. The second improvement in Recall Figures (Versions I and II combined) requires that some of the operators in a second measurement introduce a significant number of different but equally pertinent documents.

In brief, while recall improved significantly with four measurements, and even more with eight measurements, the corresponding Relevance Ratios decreased only nominally. The inverse relationship is trivial. For practical considerations, observed deterioration of the Relevance Ratio while improving Recall is entirely acceptable to the designer and manager of any system. Furthermore, statistical analysis presented in this paper has determined the significance of differences in relevance ratios.

Although we expected an improvement of the Recall Figures for statistical reasons, the size of the increase is considerably larger than we expected.

One may rationalize, therefore, that the system has a high recall "potential" which operators in this test usually do not realize. One might further argue that the use of a team in such a search is not inconceivable and would even be practicable. Moreover, several measurements must be taken in any system requiring feedback to direct the search, and two measurements with this system are not excessive. The one point that one might take issue with is the validity of the team results in view of the relatively small size of the collection, and this aspect will require further analysis.

Indications are, thus, and rather fortuitously, that the ABC system, having a high Relevance parameter, need not necessarily be content with a low Recall Figure. It therefore meets requirements stated also by other documentalists. Indeed the system can demonstrate high Relevance Ratios, while realizing high Recall Figures by using (1) teams, (2) additional measurements, and (3) perhaps greater effort by an operator, a factor we will discuss later.

ABSTRACT MODEL

In order to explain the absence of the detrimental inverse relationship between Relevance Ratio and Recall Figure when we compared individual efforts with team efforts (having particularly in mind the high rise of Recall Figures), we prepared a simple model for one imaginary series of retrieval runs.

Only one major assumption was made in constructing the model, and this was derived from our test data. Because the ABC Relevance Ratio had averaged 87.1 percent and had shown a great consistency throughout the test, we felt justified in asserting a system relevance parameter, in this case a conditional probability for a retrieved item to be relevant of $P=0.80$. The number of relevant documents in the system responsive to the imaginary request was fixed at 8, and the number of documents withdrawn in successive or independent retrieval operations was to vary from 1 to 3, 5, 9, 10, 11, 16, and 20 as shown on Fig. 1.

It is evident that x (the number of relevant documents found in each run) cannot exceed the number $r(=8)$. Relevance and Recall Ratios, therefore, develop as shown on Fig. 1.

According to our stipulation, the system operates with a 0.8 Relevance Ratio that will prevail as long as the product of $np=x$ does not exceed the value of $r(=8)$.

Once the product attains $r=8=x$, then the probability figure decreases, since x can no longer increase with larger n. Relevance and Recall Ratio therefore develop as shown on Fig. 1. The Recall Ratio rises steadily from its lowest value to the peak being reached when x equals r. In the subsequent runs this ratio remains unchanged.[7]

Because this abstract model, based upon a general characteristic of the test data, exhibits its peak relevance potential from the start independent of the efforts (as measured by n, the number of items withdrawn during the given run) that the retrieval operators exert, the following conditions prevail:

1. The Recall Ratio is proportional to the number of documents retrieved (n), and to the number of relevant documents located (x), until $np=x=r$ its optimum point; and
2. After the optimum is reached, an increase of n must result in a deterioration of the Relevance Ratio and can provide no improvement in the Recall Ratio.

From these facts we can draw the additional practical conclusions, that

1. A prerequisite for determining ABC Recall Ratio is that in addition to employing teams all individual operators make also a reasonable attempt to withdraw the properly tagged documents relevant to the request.

[7] Stephen Pollock (4) has independently reached conclusions virtually identical with those represented by the abstract model.

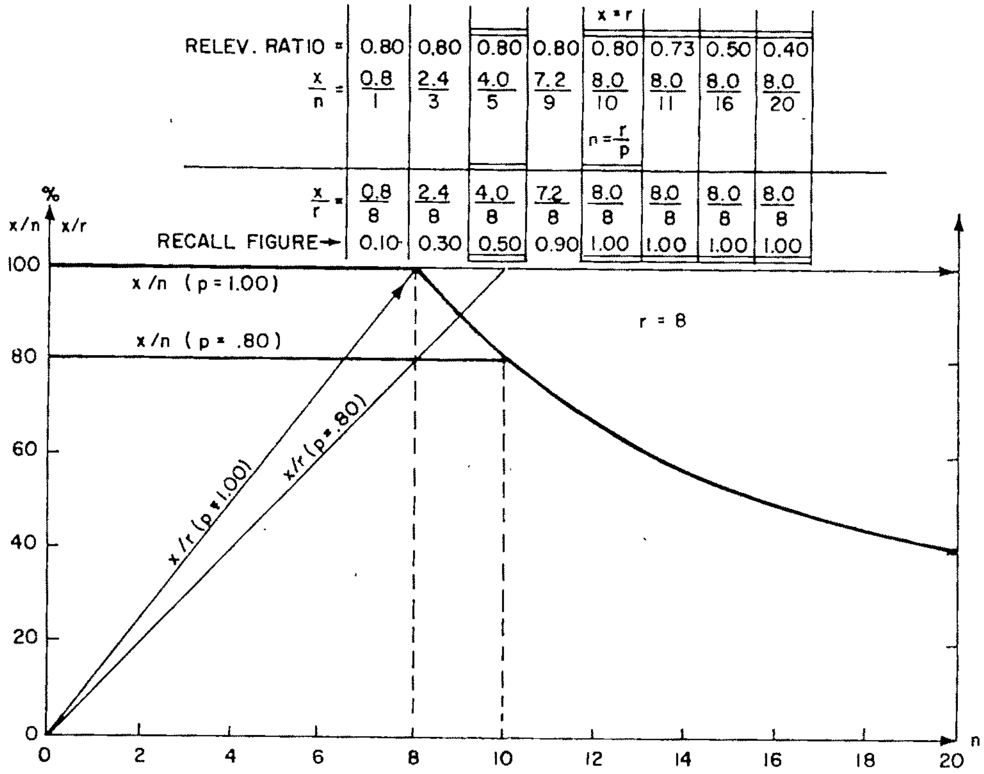


FIG. 1. Abstract model for a sequence of retrieval runs with increasing n when $r = 8$ (assumption: $p = 0.80$)

2. On the other hand, each system should have built-in features that hinder the operator from uneconomical retrieval efforts (when he reaches the $x = r$ level). In the ABC system this feature is the capability of matching concepts unambiguously.

With the abstract model we gain a better understanding of the interdependence of the factors n, x, and r and of the influence exerted in particular on the Recall Ratios observed in our test.

The model presented is no more than a conceptual structure, a generalization formed from a few isolated observations. It is an ideal type as defined by Max Weber[8] and as such may be used as a tool to interpret and compare empirical or experimental data; it is not an objective in itself. Its validity as an evaluator in this case must be established through application not only to some individual performances, but also to the over-all test.

With these requirements in mind, we return to the statistical data of our test and retabulate them for a suitable and more exhaustive analysis.

EFFECT OF PARAMETERS

In a preliminary operation, we organized the 100 document-based requests used in the test according to number of relevant documents available in the collection. The distribution of requests is shown in Fig. 2. Requests with $r = 6$, 8, 10, and 14 were then selected for analysis; and their average Relevance Ratios and Recall Figures[9] tabulated by increasing values of n to produce a structure parallel to the one of our theoretical model. Averages are recorded on Table 5.

In all the selected groups, the average Recall Figures rise with the increase of n. The explanation for this relationship is, of course, quite simple. The consistently even level of the Relevance Ratio is the *outstanding* characteristic of the ABC retrieval method not only as an

[8] Max Weber, Die objektivität sozialwissenschaftlicher und sozialpolitischer erkenntnis; in: Gesammelte aufsätze zur wissenschaftslehre, 1922, p. 174, 179, 190–191, 194. Max Weber developed his "Idealtypus" for the purpose of giving the political and social scientist, the historian and the economist a method or tool with which to identify and characterize the concrete, causal relations in social or historic life. In applying his ingenious method to the study of storage and retrieval processes we are guided by the similarity of problems, in particular the disturbing variety of manifestations the human element introduces as author, processor, evaluator, seeker, and searcher of information. Doubts have been cast on the historians' possible success in reducing observed reality to definition, organization, evaluation, and comparison; but they were dispelled. The "Idealtypus," carefully derived and formed from the observed significant facts and used as a yardstick, may lead to the same results in the discipline of documentation. However, we must keep in mind the basic difference in methodological approach. The historian employs the model or "typus" to characterize and define the singularity of a given historic fact (Th. Schieder, *Möglichkeiten und grenzen vergleichender methoden in der geschichtswissenschaft*, in Historische Zeitschrift, v. 200 [1965], pp. 529–551, in particular pp. 544–545, 550–551), while the documentalist must first determine the variations (for the better or for the worse) brought about by individuals before he can identify the performance of the system as such and objectivize its characteristics for measurement and comparison.

[9] The subsequent report (reference 8) proves the statistical validity of Recall Ratios for groups with identical r's; however, this distinction has not been made in the text to avoid a possible confusion of the reader.

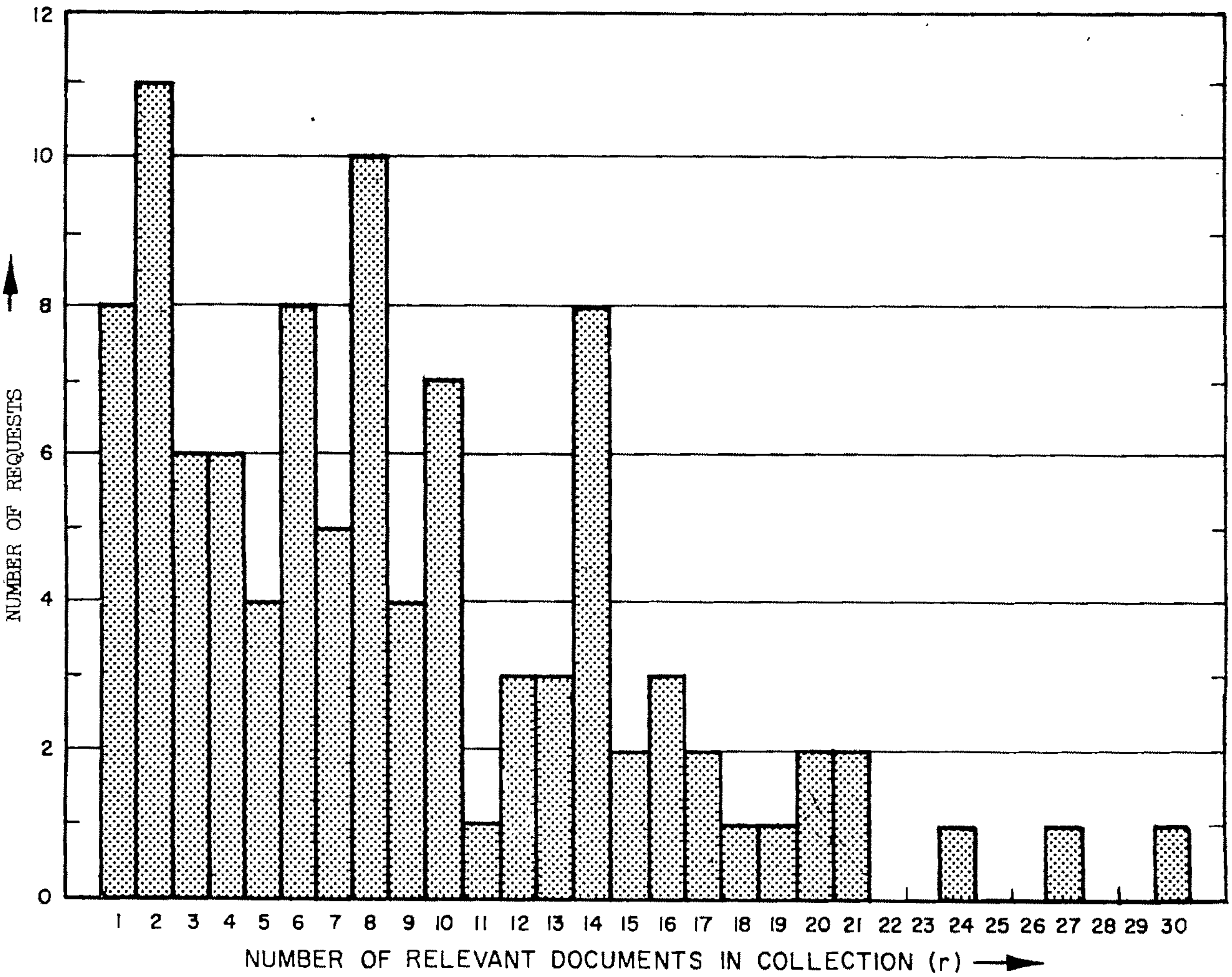


FIG. 2. Distribution of the 100 requests according to r

TABLE 5. The Influence of n on ABC Recall Figures and ABC Relevance Ratings for Four Different Values of r

	$r = 14$		$r = 10$		$r = 8$		$r = 6$	
n	Avg. Recall Figures *	Avg. Relev. Ratios	Avg. Recall Figures	Avg. Relev. Ratios	Avg. Recall Figures	Avg. Relev. Ratios	Avg. Recall Figures	Avg. Relev. Ratios
0								
1	6.0 (7.0) †	88.2	8	84	13	100	17	100
2	12.5 (14.3)	86.1	18.5	97.1	21	84	33	99
3	19.4 (21.4)	82.6	26.3	92.0	34	90.6	39	78
4	29 (28.6)	83	36	94.5	50	100	52 ‡	78
5	36 (35.7)	78.4	43	90.3	63 ‡	100	67	80.4
6			55 ‡	96.2	63	84		
7	43 (50.0)	86	52.5	78.8	88	100		
8	53.5 (57.0)	93.6	63	82.7				
9	64 (64.3)	99.6						
Average		87.2		89.5		94.1		87

* Averages were computed over the combined results for each n.

† Optimum values. The actual average value is 1.8 percent under average optimum value.

‡ These ratios correspond to the pooled recall ratio (55.5%) obtained by a team of four operators testing Version I of the ABC method with 86 freely styled requests.

over-all average, but for every r-group and with only minor deviations, every successive n-group.

The same tabulated test results make it also possible to solve a problem we have previously raised without providing satisfactory explanations. We refer to the Recall Figures obtained by team effort when groups of four using identical requests of the freely styled set increased the score to 55.4 (and 59.7) percent (Table 3, col. 4 and 8). In this particular instance, the Recall Figure found on the summary table (Table 5), under discussion amounted to 63 percent. If we make appropriate adjustments for the different r-values in the $r=14$, $r=10$, and $r=6$ groups, we arrive at the corresponding Recall Figures of 53.5, 55, and 52 percent respectively. Because all the corresponding figures prove to be functions of n within their particular (r) groups and possess nearly identical values throughout the system, it is not the participation of a larger number of persons in the search (or the larger number of measurements taken) that brought about these higher Recall Figures, but the greater number of different documents (n) withdrawn and used for the calculation.

The average Relevance Ratios and Recall Figures for runs yielding identical n/r are plotted against n/r in Fig. 3.[10] As expected, the Recall Figures increase approximately linearly with n, but become more erratic as n approaches r. The Relevance Ratio shows a slowly increasing tendency to decrease. Some of the scatter in the values as n approaches r can be explained by the relatively small samples available for averaging in this region, as is indicated by the 50 percent point. Even so, this figure illustrates the potential of the system as realized by various operators. It further illustrates the obvious relation that so long as Relevance Ratio can be maintained at a relatively constant value, and n is smaller than r, the Recall Figure is merely a function of n.

Furthermore, the consistently high Relevance Ratio fully explains why the Recall Figures computed from the returns of the $r=14$ group (Table 5, col. 1) are only 1.8 percent short of the feasible optimum results; and a brief analysis discloses that the apparently low ABC Recall Figure for the test of the entire ABC method presents actually a maximum value, too. As indicated in Table 6, the average retrieval run yielded only 1.8 relevant documents out of an average of 8.5 relevant titles in the collection. The ABC recall figure could or should therefore not have been larger than (1.8)/(8.5) or 21.2 percent.

The observed test data thus approximate those of the abstract model presented on Fig. 1. The first-generation ABC system exhibits persistently high Relevance Ratios independent of n. The number of documents (for $n<r$) withdrawn, but its Recall Figures are pre-eminently determined by the size of n.

What remains is a determination of the causes that were responsible for the failure to obtain a greater (average) number of documents from the collection during the test. We base our analysis on the scores made by different operators with identical requests.

As a representative sample, we have compiled all the

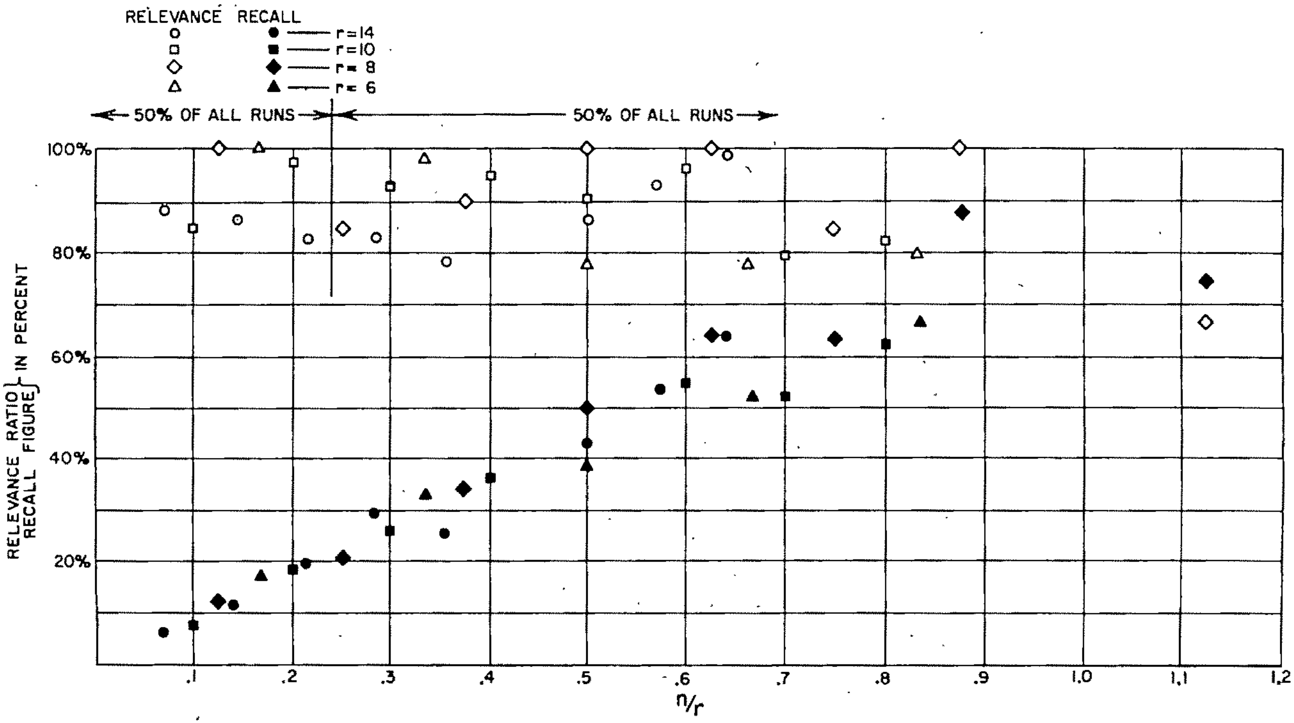


FIG. 3. The influence of n on ABC Recall Figures and ABC Relevance Ratios for four different values of r

[10] A more complete presentation of the Recall Ratios (so termed because of the limitation to distinct r groups) obtained by various runs for various r's is shown in Fig. 4. The lines indicate the optimum Recall Ratio for the respective r's.

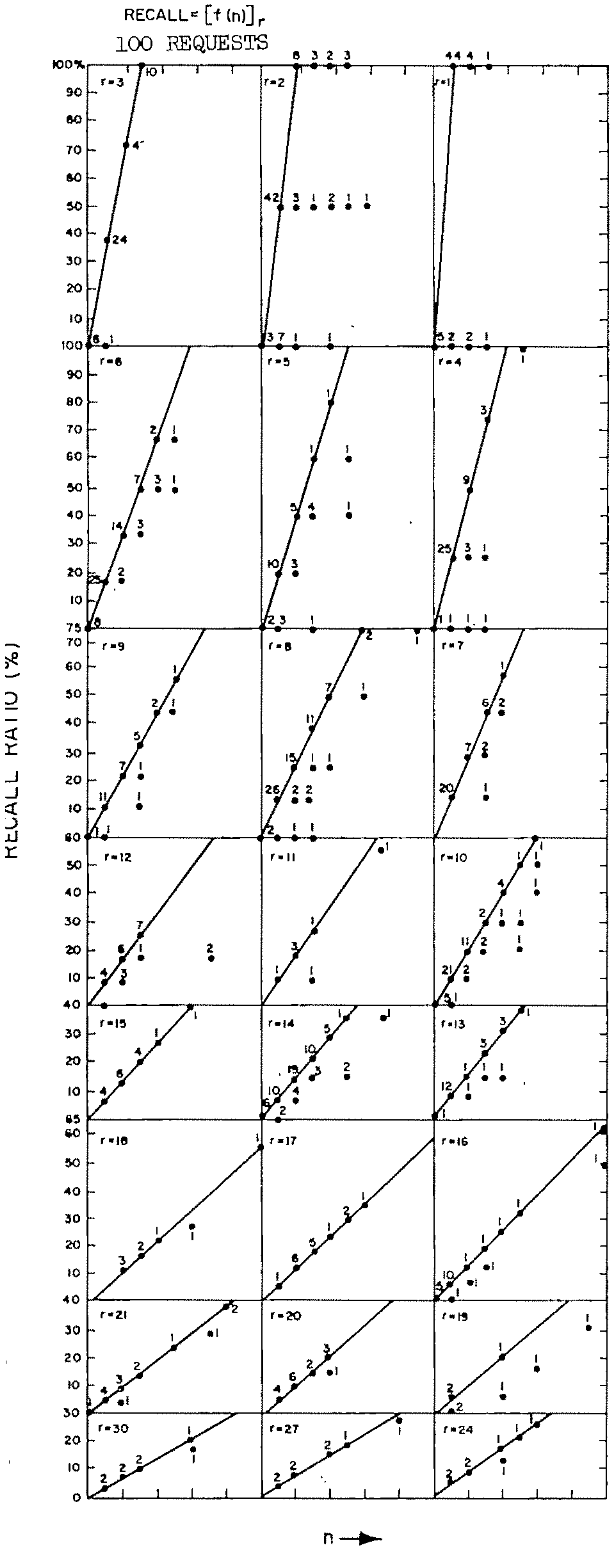


Fig. 4. Recall ratios vs. number of retrieved documents for requests with identical number of pertinent documents in collection (lines indicate optimum recall ratio)

Table 6. Test Results of ABC Methods I and II Combined

Sets of requests	x_1*	x_2†	n_1‡	n_2§	r‖
36	2.34	1.99	2.64	2.25	8.18
100	1.85	1.70	2.15	1.99	8.60
136	1.97	1.8	2.28	2.05	8.50

$$\text{Average relevance ratio: } \frac{x_1}{n_1} = \frac{1.97}{2.28} \cong 86.4 \text{ percent}$$

$$\text{Average recall ratio: } \frac{x_2}{\bar{r}} = \frac{1.8}{8.5} \cong 21.2 \text{ percent}$$

* x_1 = Average number of relevant documents retrieved (trials without results not counted).

† x_2 = Average number of relevant documents retrieved (trials without results are counted).

‡ n_1 = Average number of documents retrieved (trials without results not counted).

§ n_2 = Average number of documents retrieved (trials without results not counted).

‖ r = Average number of relevant documents in the collection.

observed ABC Recall Figures for 14 randomly selected requests: for seven requests each being related to 15 to 24 relevant documents in the collection; and for an additional seven requests satisfied by only one to four relevant titles (see Table 7). The values are listed by the rising number (n) of documents selected and withdrawn in the particular run.

A brief study of the table makes it evident that while a few operators utilized the capability of the system more extensively (by withdrawing for example up to 10 documents for requests with a bigger number of corresponding relevant documents in the collection; and up to five documents for the requests addressed to a small number of presumably related papers), the majority of the participants considered their task completed when one to two documents (concerned with the request) had been identified. Within this limitation intentionally or subconsciously placed on their contributions, the volunteers performed a very good job. In fact their quantitative output increased in proportion to the decline of r (= number of documents relevant to the particular request). This can be evidenced by the analysis of the results from 60 valid retrieval operations performed only with requests for which no more than one single document provided the appropriate or acceptable response (see Table 8). In 50 out of these 60 events, the one available document was recovered, and a Recall Figure of 83 percent was obtained.

• Discussion

The analysis of the data prevents us from characterizing the ABC system by a low Recall Figure. On the contrary, we have shown that it has a relatively high recall capability, and we have in addition shown at least one means of achieving this consistently.

Also, we are in no way disposed to attribute low Recall Figures to a deficiency in the operators alone. The majority were scientists and engineers who had no previ-

Table 7. Average Recall Figures Calculated from Multiple Responses to 14 (high and low r) Requests and Arranged by Increasing n (in percent)

Requests no.	r	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$
135	1		100	100	0							
63	2		50	100		75	100					
50	2		0		75	100	75					
133	3		33	67	100							
42	3	0	27									
84	4		25	50	75							
45	4		25	42			100					
76	15		7	13	20	27		40				
74	16		6	9		25	30					60
78	16	0	4.5		13							50
57	17		6	12	17	24	30					
60	21		5	10	14		24			38		
53	21	0	3.8	7.5				30				
80	24		4	8		15	20	25				

ous familiarity with the system. They volunteered their services but were not relieved from regular responsibilities. An early return to the work bench may have been on the mind of many. At least, such a desire on their part would be fully understandable. Moreover, the test proceeded under realistic conditions, where the operators were merely instructed in the use of tools; they were at no time urged to exhaust all available avenues and to pursue all cross references offered by the system, so that Recall Figures would be as high as possible. Team results and individual performances indicate that such could have been done without a significant deterioration in Relevance Ratios.

Other contributing factors to the average low Recall Figures as was discussed in the preceding report (2, pp. 26–29) were certain disadvantages of the first ABC format.

The design of the second-generation model and in particular the clearer and more appropriate organization of the descriptive sentences in the new ABC dictionary as well as the addition of a filter system will facilitate the faster acquisition of related subject matter and assure the withdrawal of a larger number of documents, which should raise the Recall Figures.

Although we have shown that the ABC system can yield high Recall Figures and although we have made certain provisions to improve the probability that larger figures are obtained, analysis leads us to the conclusion that Recall is not an absolute measure of system performance. With the same set of questions applied to the same collection the average r will vary from investigator to investigator because of different backgrounds and requirements, and obviously, the average r will vary from one set of requests to another set of requests.

Table 8. Retrievals Resulted from Requests with $r=1$ *
(Among the 100 Document-Based Requests)

Question no.	Short dict.								Long dict.								Average scores	
	1a		1b		2		3		1a		1b		2		3			
	n	x	n	x	n	x	n	x	n	x	n	x	n	x	n	x	Recall	Relevance
43	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	100
81	1	1	1	0	1	1	–†	–	0	0	1	1	1	1	0	0	57	80
99	3	1	2	1	1	1	1	1	2	1	1	1	1	1	1	1	100	67
121	1	1	1	0	1	1	1	1	1	1	3	1	1	1	1	1	88	70
125	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	100
129	1	1	1	1	2	0	–	–	1	1	1	1	2	0	1	1	71	56
135	2	1	2	1	1	1	1	1	–	–	1	1	3	0	1	1	86	55
136	1	1	1	1	0	0	0	0	1	1	1	1	0	0	–	–	57	80
																Averages:	82.4%	66%

In 44 out of 60 valid runs, it was found that $x=1$, and $n=1$; that is,
Recall = 100% and Relevance = 100%.

* There are 8 Requests with $r=1$, with 60 Valid Retrieval Runs.
† – Means no data for evaluation available.

With the limited average effort invested in the test, it is obvious that the Recall Figures must decrease with increasing r. According to the returns of the test the Recall Figures dropped from 82.4 percent for the document-based requests with $r=1$ (Table 5) to 46.6 percent for those with $r=2$. All these considerations lead to the conclusion that Recall Figures of our test depend on a great number of variables and subjective factors and appear to be a very debatable measuring rod for ABC Recall capability.

The frequently advanced theory of an inverse relationship between Relevance Ratio and Recall Ratio, which can be alleviated only by clumsy and most uneconomical operations cannot be substantiated by our organized data. Although individual operators were able to attain high Recall Figures, their Relevance Ratios in some cases were also higher, and only seldom were they lower. Although higher Recall Figures are obtained in a team approach, Relevance Ratios are not significantly lowered. Reduced Relevance Ratios with the KWIC title list were accompanied by no increase in Recall Figures (Tables 3 and 4) and are furthermore indicative of the quality of titles when compared with analytic descriptions.

If the tested ABC system can be described by a model, the logical choice is one where we have a high probability that a retrieved document is relevant, and the recall figures are pn/r. Whether n can be raised by individual operators at will to $pn/r \approx 1$ without significant decrease in Relevance Ratio cannot be answered without a test of the completed second-generation model. Indications from this test (team results and individual scores) are that if data are re-arranged according to increasing effort measured by n, the Recall Figures have been greatly increased without a conspicuous deterioration of the Relevance Ratios.

• Conclusion

According to the established procedures of this test, identical requests were used repeatedly, first of all, to expose the system to a cross section of the user population. As a rule, different operators tested the ABC method eight times with each of 136 standardized requests. The multiple test produced a greater confidence in the average Relevance Ratios because of the consistent scores. It brought to light the inherent deficiency of Recall as a measure of ABC system performance. It permitted, through a comparison of the different individual responses to identical requests, the separation of characteristics of the storage and retrieval system from the attitudes or efforts of the operators. This allowed a more adequate evaluation of the system as such.

The multiple testing method made it also possible to determine the degree of bias that might be introduced by the program and the procedures of the test. For example, the relative independence of the two runs by identical operators using the same request is corroborated by the team Recall Figures. The combined scores of the eight retrieval operations (Tables 2 and 3) had resulted from contributions of only four individuals who had tested two versions of the ABC Dictionary. Because of their success in raising the Recall Figure from 55 to 78 percent through the second run, it is evident that the same operators located different, but equally relevant materials, and therefore remained relatively unaffected by the findings during their first performance.

The request of our statisticians for a sizable test collection has been apparently vindicated by the results. While a small collection may be useful in developing a mathematical model prior to the design of a test, or in testing a fully automated system, the value of the ABC system and its components, including the average contribution of the analysts (or subject cataloguers), could not have been examined by a representative sample of the user population and by a representative set of requests. In other words, the test of a small collection would have made it difficult to predict the operational capability or utility of the system within a large and rapidly growing real collection.

Test results show that the ABC system has a consistently high Relevance Ratio. During the test, the average Recall Figures were relatively low. However, analysis shows that the ABC recall capability is high (in the majority of cases observed average Recall Figures were less than 10 percent below the optimum obtainable for given n and r), and that procedures can provide for both high relevance and recall.

• Future Work

Instead of pointing to the relatively good and consistent Relevance Ratios, we wish to direct our future attention to the consistent amount by which the system errs (14.9 percent of the documents withdrawn are not relevant). What are the contributing factors to this overall deficiency: (a) the inability of the searcher to understand correctly the descriptive sentences in the ABC dictionaries, (b) the lack of appropriate words or phrases used by the analysts in their cataloging, or (c) still other elements which require correction of the system?

Other subjects which will require discussion and study are: the time factor, the influence of operator group background on the scores, creation of mathematical models, the analysis of phrases and sentences for comparisons and standardization; preparation of SOP's for the analysts, the automatic production of thesauri, evaluation of the KWIC title approach, the development of new testing methods and new yardsticks for the evaluation of systems, and the comparison of the results with the results of different tests.

General Conclusions

Test results published to date suggest that coordinate index systems are faced with the alternatives of either "vastly extending the power of the basic coordinate indexing process or . . . replacing this process by an altogether different one" (5).

Without exceptions designers, operators, and evaluators of coordinate-index type retrieval systems experienced and acknowledged the unavoidable deficiency of the inverse Relevance-Recall relationship.

Highly sophisticated methods and computer programs as well as computers with giant memories are the prerequisites[11] for future information systems that will place the scientist in a position where he (starting with a given set of terms or a preliminary formulation of his problem) can guide a retrieval operation to a successful completion in a personal direct dialogue with the computer.

While such solutions cannot be realized before two, and probably many more, decades have lapsed, the ABC system approximates the performance of such systems today; it can easily be adjusted for exhaustive manual retrieval acceptable to the scientist with respect to flexibility, quality, and speed of output; it can also, without large expenditures in time and dollars, provide: (a) fully automated retrieval runs on the basis of Boolean combinations and probably the application of a vector method, and (b) mechanized production of a dictionary or thesaurus. Future research will eventually lead to the automatic standardization of terminology and syntax as well as to mechanized semantic retrieval.

[11] The SMART system with its multiple and consistent approaches is an outstanding example for a program moving toward the final objective (6, 7).

References

1. Altmann, B., A Natural Language Storage and Retrieval (ABC) Method: Its Rationale, Operation, and Further Development Program, *Journal of Chem. Documentation*, 6:154–157 (1966).
2. Altmann, B., *A Multiple Test of the ABC Method and the Development of a Second Generation Model, Part 1, Preliminary Discussions of Methodology*, TR-1295, Harry Diamond Laboratories (April 1965).
3. Menden, W. H., and B. Levy, *A Multiple Test of the ABC Method and the Development of a Second Generation Model, Part 3*, TR-1334, Harry Diamond Laboratories (in preparation).
4. Pollock, Stephen, *The Normalized "Sliding" Ratio Measure* (Technical Note CACL No. 19), Arthur D. Little, Cambridge, Mass., July 7, 1965.
5. Rial, J. F., *Final Report on the ROUT (Retrieval of Unformatted Text) Document Retrieval System* [Contract AF 19 (628)-2390], (ESD-TDR-64-96; TM-3869), Mitre Corp., Bedford, Mass., pp. 63–65, May 1964.
6. Salton, G., and M. E. Lesk, The SMART Automatic Document Retrieval System—An Illustration, *Association for Computing Machinery Communications*, 8 (No. 6): 391–398 (1965).
7. Salton, G. (Project Director), *Information Storage and Retrieval, Scientific Reports . . . to the National Science Foundation*, No. ISR-1 seq., Harvard University, Computation Lab., Cambridge, Mass. (beginning with June 1966, Cornell University, Department of Computer Science).

Brief Communications

Bradford's Law and the Keenan-Atherton Data

Bradford's methods are applied to the Keenan-Atherton data. The results do not fit Bradford's Law.

Bradford's Law (1) states: "If scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the number of periodicals in the plot, and the succeeding zones as segments of the remaining straight portion of the plot with each zone having the same number of references as the nucleus. The curve for

Bradford's curve is drawn by plotting the running sums of references against the logarithms of the running sums of titles. The nucleus is defined as the initial curved part of the nucleus and the succeeding zones will be as 1:n:n² . . ." the zones has to be a straight line if the ratio 1:n:n² . . . is to hold.

The Keenan-Atherton data (2) was plotted in Bradford's manner, and is shown as Fig. 1. If we take Bradford's criterion for the nucleus as the curved section of the plot, and his condition that the succeeding zones contain the same number of references as the nucleus, we get a fair approximation by setting the end of the nucleus at 6,581 references, the boundary between the first and second zones at 13,456 references. The terminal naturally falls at the total number of references, 20,287.

The curve does not, however, support Bradford's ratio 1:n:n² . . .; to have done so, it would have to have taken the dotted extension shown on the figure. The reason for the deviation lies in the 10 percent higher number of titles having the minimum number of references in the Keenan-Atherton data. This is shown in Table 1.

The Keenan-Atherton study would therefore seem to indicate that Bradford underestimated the percentage of titles having a minimum number of references, and therefore drew an invalid ratio.

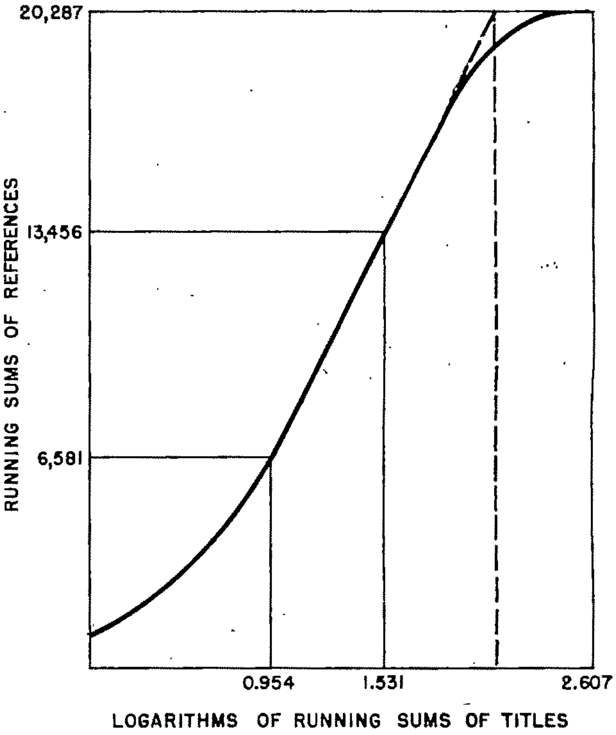


Fig. 1. Bradford curve for the Keenan-Atherton data

References

1. Bradford, S. C., *Documentation*, Public Affairs, Washington, pp. 110–120, 1950.
2. Keenan, S., and P. Atherton, *Journal Literature of Physics*, American Institute of Physics, New York, 1964.

Ole V. Groos
AFCRL Research Library
Bedford, Massachusetts

Table 1. Distribution of titles and references between nucleus and two zones.

	Bradford geophysics titles	Bradford lubrication titles	Keenan-Atherton titles	Bradford geophysics references	Bradford lubrication references	Keenan-Atherton references
Nucleus	9(2.8)	8(4.9)	9(2.2)	429(32.2)	110(27.8)	6,581(32.4)
Zone 1	59(18.1)	29(17.7)	25(6.2)	499(37.5)	133(33.7)	6,875(33.9)
Zone 2	258(79.1)	127(77.4)	371(91.6)	404(30.3)	152(38.5)	6,831(33.7)

Solution of Boolean Equations Through Use of Term Weights to the Base Two

In a recent communication by Brandhorst (1), a method of weighting is presented for the solution of Boolean equations. In the paper, Brandhorst notes that ". . . though term weighting had its advantages, nevertheless there were some equations that could not be reduced in this way."

There is an alternative to the type of term-weighting that Brandhorst presented. While not satisfying a simple (single) weight-limit, the technique does afford a unique solution to a Boolean equation when used in conjunction with a table-lookup following assessment of the truth-value of each term in the set.

By assigning a series of weights, $2^0, 2^1, 2^2, \ldots, 2^{n-1}$ (for $n =$ number of elements in the set), to each element in the set, a unique sum is guaranteed regardless of which subset is found true. A simple table-lookup following the summation with the attained sum as the argument can then locate a branch to an appropriate part of the program. If the sums (equivalent to "truth" in the equation) are few, comparisons against the obtained sum may be faster than table-lookup procedures.

For example, a set of four elements, A, B, C, D, has a maximum of $2^4 = 16$ possible combinations. Assigning weights of $A = 1$, $B = 2$, $C = 4$, $D = 8$ will yield a set of sums taking on all values from 0 to 15 with each sum uniquely identifying its own subset.

The method is illustrated using the same equations, in the same order, as those presented by Brandhorst. The weights of 1, 2, 4, 8 applied to A, B, C, and D are constant throughout all equations. The conditions for truth are presented as the sum (S) immediately to the right of each equation.

(1) $A(B + C + D)$ — $S = 3, S = 5, S = 7, S = 9,$ $S = 11, S = 13, S = 15.$

(2) $A + B + C + D$ — $S \neq 0.$

(3) $A \cdot B \cdot C \cdot D$ — $S = 15.$

(4) $A + (B \cdot C \cdot D)$ — $S = 1, S = 14, S = 15.$

(5) $(A + B) + (C \cdot D)$ — $S \neq 0, S \neq 4, S \neq 8.$

(6) $(A + B) \cdot (C \cdot D)$ — $S \geq 13.$

(7) $(A + B) \cdot (C + D)$ — $S \geq 5$ where $S \neq 8$ and $S \neq 12.$

(8) $(A \cdot B) + (C \cdot D)$ — $S = 3, S = 7, S \geq 11.$

The method for evaluation of the resultant S must, of course, depend upon the complexity of the range of truth values; equation (1) is obviously best evaluated by a table-lookup while equations (2), (3), and (6) can be handled by simple comparisons.

This technique is often used in computer programming where, for example, any of 2^n different subroutines (branches) must be taken depending on the truth subset of n elements. Use of a continuously bifurcating tree-structure approach can require as many as 2^n tests in order to locate the appropriate branch. The weighting system described here will require n tests, n summations (if all elements are true), and a table-lookup.

References

1. Brandhorst, R. T., Simulation of Boolean Logic Constraints Through the Use of Term Weights, *American Documentation*, 17 (No. 3): 145–46 (1966).

Howard P. Iker
University of Rochester
School of Medicine and Dentistry
Rochester, New York

A Decentralized National Chemical Information System

Many witnesses of the chemical information scene appear to agree that the inevitable National Chemical Information System is doomed to a short life if the hard copy of all the literature must be stored in a geographically single, monolithic structure.

The proposal that follows is an alternative. There should be a minimum of centralization of the document collections. Use the current university collections, supplemented by the Center for Research Libraries, the CAS library, and others.

During the past decade a large amount of discussion and a significant amount of work have centered around the construction of bibliographic and search techniques, both theoretical and practical, both experimental and operating, with and without regard to cost. All of the problems have not been solved for handling structures of chemicals, for searching concepts, for formulating questions, etc. Nonetheless, the burden of interest has largely neglected the final step of presenting the requestor with hard copy, accurately, promptly, and cheaply.

We suggested a solution obliquely in 1961 (1). The concept of a National Chemical Information Center was not being discussed in 1961. The technical information problem had not then been called to the attention of the President of the United States. There was no COSATI. The Chemical Abstracts Service had only begun its research program. Few libraries had computer programs. But the need for the presentation of hard copy, the journal article, to the requestor was at that time, and still is, the bottleneck in the transmission of technical information (2).

All of the large chemical libraries in universities in the country could be linked by wire service to the major searching services. When a search has produced its output of references, the library nearest the requestor will be notified of the references needed immediately, and for a small fee will photocopy them, and mail them promptly to the requestor. Alternately, the requestor could order needed references directly from an identified, certain source and expect to receive them by return mail.

Every known document which conveys new chemical information would be catalogued and available to this wire service. If some of the 10,000 journals covered by *Chemical Abstracts* were not subscribed to at all in the United States, a small library could be formed at the site of the searching service to guarantee that all original articles in chemistry, including the "obscure," were available.

We recognize that the economics of interlibrary loans and photocopies are currently not in favor of the lending or copying library. We recognize also that there are inefficiencies in university libraries today that could be corrected with an initial capital outlay for the long term benefit. A Federal subsidy to these university libraries would initiate the program and would be continued only so long as it functioned as an incentive to the libraries' administrations to improve their efficiencies through innovation.

Those libraries failing to justify continued subsidy would be dropped from the National Chemical Information System, and other university collections would be sought to fill the need of supplying copies of any articles that have appeared in the chemical literature.

An example of how this concept might work is as follows:

Some Federal agency, such as the National Science Foundation, might contract with the Chemical Abstracts Service to be responsible for searching the world's chemical literature, with the aid of computers if necessary, and for maintaining administrative control of the wire service system to an adequate number of university libraries throughout the fifty states to guarantee:

1. That a copy of every article is available, no matter how obscure.
2. That through the sale of tokens at a fixed price, the out-of-pocket costs for the photocopy service would be borne by all requestors, whether U. S. government, foreign, nonprofit, or commercial.
3. That prompt service be maintained for requestors.
4. That there would be a continuing competition among the chemical libraries of the country to qualify as a member of the National Chemical Information System and for its subsidy.

References

1. WILKINSON, W. A., and W. H. WALDO, New Information Services—A Practical Approach, *Jour. Chem. Doc.*, 2: 175 (1962).
2. NICHOLSON, N. N., Service to Industry and Research Parks by College and University Libraries, *Library Trends*, 14: 262 (1966).

W. A. WILKINSON and W. H. WALDO
Monsanto Company
St. Louis, Missouri

Letters to the Editor

Dear Sir:

Concerning the Brief Communication by W. T. Brandhorst, "Simulation of Boolean Logic Constraints Through the Use of Term Weights" [*American Documentation*, July 1966], the "newly realized relationship" between Boolean expressions and term weights has been known for quite a few years and has been the subject of intensive work for the past five years in the field of switching theory.

The "weights" and "weight limit" of the Communication correspond to the "weight" and "threshold" of threshold logic devices; the switching theorist calls Boolean functions that can be synthesized by a single threshold device "linearly separable functions." The "group weight" procedure described in the Communication corresponds to the use of multiple threshold devices for the synthesis of functions which are not linearly separable.

Many of the papers in the field appear in the Institute of Electrical and Electronic Engineers (IEEE) Transactions on Electronic Computers, and are concerned with synthesizing Boolean expressions in terms of the fewest number of threshold devices. The topic is also covered in recent textbooks on switching theory.

References

1. Hopcroft, J., and R. Mattson, Synthesis of Minimal Threshold Logic Networks, *IEEE Trans. on Electronic Computers*, EC-14: 552–560 (Aug. 1965).
2. Winder, R., Single Stage Threshold Logic, *Switching Theory and Logical Design*, AIEE Publication S-134, pp. 321–332 (Oct. 1960).

O. Firschein and M. Fischler
Electronic Sciences Laboratory
Lockheed Missiles & Space Company
Palo Alto, California

Dear Sir:

My first reaction to the letter from Messrs. Firschein and Fischler was one of personal embarrassment that I should have been unaware of the work they described. However, upon examining the references cited by them I realized that we apparently have in this case an excellent example of a failure of information transfer between two areas of endeavor.

The important thing is not that the relationship between logical expressions and weighted expressions was known and used in other fields such as switching theory, but that it was *not*, as far as I could ascertain, being applied anywhere in document-retrieval efforts. There was no claim that the basic concept was totally new, but that it did not appear to have been explicitly realized in the information-technology field and that this was quite surprising considering the length of time that field has been using Boolean Logic in its search strategies.

When we first hit on the concept described in my Brief Communication, which appeared in the July 1966 issue of *American Documentation*, I did a literature search in an attempt to discover what other workers in the information-retrieval area might be using the same technique. I didn't find any. Nor did I find any papers at all on the subject in *American Documentation, Special Libraries*, or any other journals in the documentation field. Textbooks such as Becker & Hayes' *Information Storage and Retrieval: Tools, Elements, Theories* likewise were silent on the subject. I sent the paper around to several workers in the field and all indicated the idea was new to them.

Apparently few of us do any reading in the field of switching theory or are able to extrapolate from the terms and applications of that field to our own. Now two gentlemen from that field point out that synthesizing Boolean expressions in terms of the fewest number of threshold devices is a topic of long-standing concern with them and hardly new.

What I would now like to know is just when did awareness of this relationship enter the field of documentation? My literature search may not have been thorough enough. It was obviously too restricted in scope. Perhaps other workers are making use of the concept. (That is one reason it was made a Brief Communication instead of an article.) I would like to hear from anyone who might have information on this subject. If, however, our use of the concept should represent its first explicit appearance in this field, then we have an interesting time lag to explain.

W. T. Brandhorst
Documentation Incorporated
Bethesda, Maryland

Dear Sir:

In his Letter to the Editor in the July 1966 *American Documentation* (p. 148), Mr. Robert Jordan argues for the use of full forenames, rather than merely initials, in citation indexes and other large personal author listings, e.g., Robert Thayer Jordan rather than R. T. Jordan.

The basic justification he gives for full forenames is that the practice would permit users to discriminate between individuals having common last names and first names beginning with the same letters, e.g., "Richard Jones," "Robert Jones," and "Raymond Jones" are not all reduced to "R. Jones."

This argument assumes, however, that in every appearance of a particular author's name, one encounters the same pattern. In the field of the technical report literature such an assumption is decidedly not valid. Whether because of the "corporate" nature of their production or the welter of sign-offs these documents frequently go through, there is in actual practice little consistency in the way given personal names appear on technical reports.

What this means is that if forenames were extracted *as found* in actual technical reports, and repeated in large indexes, there would be a tendency to separate indexing entries for the same author. Let us take, for example, the name Robert Thayer Jones. By overstating our case slightly we might postulate the following occurrences:

R. T. Jones
R. Jones
Robert T. Jones
Robert Thayer Jones
R. Thayer Jones
Robert Jones
(Not to mention Thayer Jones, T. Jones, Bob Jones)

When you have a lot of Jones's, an "R. Jones" can file a great distance from a "Robert Jones." The use of initials alone in the preceding example would reduce six file points to two: "R. Jones" and "R. T. Jones." The use of initials is therefore one way of combating the inconsistencies met with in the report literature and providing the searcher with fewer places to look. This of course, must be weighed against Mr. Jordan's objection that it runs together all the "R. Jones" without discriminating the *Robert's* from the *Richard's*. This is true and constitutes a valid objection. In our own system we have found that the improved searchability of the file overrides the above objection and that one can usually tell from the subject matter of the report or the various sub-entry discriminators, such as the report number of the corporate source, whether the "R. T. Jones" index entry being consulted is the "Robert Thayer Jones" of interest.

W. T. Brandhorst
Documentation Incorporated
Bethesda, Maryland

Book Reviews

1/67-1R **On Retrieval System Theory.** 2d edition. 1965. B. C. Vickery. Butterworth, Washington, D. C. 191 pp.

On Retrieval System Theory, 1961, represented the first serious attempt at a reasonably comprehensive overview of the field of information storage and retrieval. It was much needed work, invaluable to students in schools of library and information science and also to the many individuals entering documentation from other areas of endeavor. In reviewing Vickery's book (*Journal of Documentation,* December 1961), A. R. Meetham stated: "a newcomer . . . would have been saved six months, spent mainly on the retrieval of information about retrieval, if the book had been ready earlier. No one else now needs to enter the field through quite such a prickly hedge."

The four years elapsing between the first and second editions of *On Retrieval System Theory* saw the publication of other texts on the subject, notably by Becker and Hayes (Wiley, 1963), Bourne (Wiley, 1963), Kent (Interscience, 1962), Sharp (London House, 1965), and Williams (Business Press, 1965). The second edition of Vickery can now be reviewed in relation to other attempted surveys of the field.

Robert Fairthorne, writing in 1958, stated that "indexing is the basic problem, as well as the costliest bottleneck of information retrieval." Seemingly, however, the importance of the indexing operation (which, in the broad sense, encompasses the surrogation of documents and of requests and the creation of a search file to allow the matching of document surrogates against request surrogates) has been overlooked in many quarters. The most publicized and most cited texts on the subject (those of Becker and Hayes and of Bourne) are largely concerned with file organization and with methods of physically implementing a retrieval system. They pay scant attention to the factors that importantly affect all retrieval systems, whether precoordinate or postcoordinate, manual or mechanized, namely: the size of the document classes defined by the index language, the extent to which document subject matter is recognized in indexing and is translated into the language of the system, and the strategies by which requests are matched against the file of document surrogates.

Vickery's text, fortunately, is concerned with "the general principles of design and operation of systems for the selection of documents containing information." It deals largely with factors basic to all retrieval systems: the surrogation of documents and requests; the structure and characteristics of index languages and the devices they incorporate to broaden or restrict class definition; file organization and coding; searching strategies; and performance evaluation. The only chapter devoted to mechanization per se discusses in a general way the extent to which equipment can be applied to the various stages of the total storage and retrieval process.

The second edition follows closely the outline of the first, although virtually all chapters have undergone considerable revision and updating. In particular, Vickery has taken full account of experimental work carried out in the area of automatic surrogation of documents. The text and bibliographies reflect a considerable amount of study and synthesis of the literature of documentation.

From the point of view of the student, using Vickery as a basic text, the organization of the work could undoubtedly be improved. In particular, Chapter 4, "Descriptor Languages," which jumps back and forth between precoordinate and postcoordinate systems in its discussion of devices used to broaden class definition (and thus improve recall) or restrict class definition (and thus improve precision) might well prove confusing to the reader not thoroughly familiar with the evolution of modern retrieval systems. Moreover, some of Vickery's statements are not strictly accurate. For example, a conventional card catalog is not an item-entry system, even though a single catalog card, containing full subject tracings, may be regarded as a unit record for a document. Rather, a card catalog, arranged by classification scheme or alphabetical subject headings, is a term-entry system. We make our first approach to such a system by going directly to the terms or class labels that best represent the subject matter we are seeking. We do not search such a file item by item except as an additional screening process within the document classes we have chosen to consult initially. In other words, we search a card catalog by consulting selected term columns in Vickery's item-term matrix, not by scanning the item rows of the matrix.

On Retrieval System Theory is not a book for the gadgeteer. Nevertheless, it remains the best available text on the intellectual aspects (of indexing, index language, and searching strategies) fundamental to the design and operation of a successful retrieval system.

F. W. Lancaster
Information Systems Evaluator
National Library of Medicine

1/67-2R **L'Automatisation des Récherches Documentaires: Un Modèle Général—le Syntol.** 1964. R. C. Cros, J. C. Gardin, and F. Lévy. Gauthier-Villars, Paris. 260 pp.

This is a report of several years work on the part of these French collaborators to formulate a new development in automatic documentation. By definition, the system is bound neither to a restrained scientific area, nor to a unique type of document analysis, nor to a single format. The system, called "Syntol" or "Syntagmatic organization language," includes a group of logical and linguistic rules which permit the retrieval of information and facilitate its manipulation by means of electronic calculators. There is a discussion, in English, of Syntol in the Rutgers series on Intellectual Systems for the Organization of Information.

Syntol is therefore an artificial language, and it is discussed in this sense. Gardin wrote all chapters except Chapter 4 and Chapter 5, which were written by Cros and Lévy respectively.

The word "syntagme" is one which is taken to include not only "key words" or "descriptors" but also their structural relationships. A dual organization is established for these descriptors. Generic relationships are specified by a "paradigmatic" arrangement, while nongeneric relationships are provided by the "syntagmatic" arrangement.

A piece of information is shown graphically with lines, with circles, and frequently with arrows. It is possible to compare graphs in such a way that varying quantities and types of information can be retrieved.

While the system may be complex, making the first reading difficult for those with a meager background in mathematics, the book is well written, and the reading, rewarding.

Pauline M. Vaillancourt
Memorial Sloan-Kettering
Cancer Center

1/67-3R **L'Organization de la Documentation Scientifique.** 1964. J. C. Gardin, E. de Grolier and F. Levéry. Gauthier-Villars, Paris. 269 pp.

In 1962 the Centre National de la Recherche Scientifique Française set up an award of 10,000 French francs, intended to bring forth new ideas in the field of scientific documentation. The rules governing the "Grand Prix de la documentation scientifique" are abstracted and presented in this volume.

The book consists of three papers whose authors were selected by a jury to share the award equally. The papers are quite different in approach, as can be expected from these authors, well-known to documentalists, who have diversified orientations and primary interests.

J. C. Gardin presents the most complete report to that date on his "Syntol," or "Syntagmatic organization language," for which he is best known. He considers all aspects of a National Center of Scientific Documentation, and, though his chief interest is in Syntol and the manipulation of information for retrieval, he nevertheless touches on the financial and organizational problems involved in the establishment of such a center on a national level. This essay follows the publication of "L'Automatisation des Récherches Documentaires."

E. de Grolier, in collaboration with Calvin Mooers, gives a general review of a new approach to organization of a large-scale center on scientific documentation as compared with using the existing diversified approaches throughout the world. The author concedes that his presentation is perhaps more schematic than is desirable, but he insists that it is best to begin anew rather than to adapt from any documentation centers now in existence and thus be trapped into solving their problems.

Several sections are proposed for this center: a depository for rarely used material; a section for commercial reproduction of documents for distribution; a center for coordinating and distributing abstracts and indexes; a center for automatic treatment of information; a center for information about scientific research "in progress" as well as published research; and a center for the publication of bibliographies automatically by progressively better computer techniques.

F. Levéry, an engineer with IBM, France, treats in some detail the mechanical aspects of documentation and of the formulation of a thesaurus.

Generally, this book is of interest to those who are involved in planning documentation centers. The article by de Grolier especially shows an awareness of the related literature, sometimes to prove a point, sometimes to cite the item to which the author takes exception.

PAULINE M. VAILLANCOURT
Memorial Sloan-Kettering Cancer Center

1/67-4R **The Politics of Research.** 1966. Richard J. Barber. The Public Affairs Press, Washington, D. C. 167 pp.

This book is an exposé of the politics of research in the United States. To quote university-professor Barber: "Although Research and Development now involves annual expenditures of about $21 billion of which more than two-thirds comes from federal funds, science retains such an aura of mystery that the scientific community has been free from the close scrutiny and skeptical appraisal that we typically regard as characteristic of the American political process." Barber proceeds to describe the magnitude and characteristics of research in this country, primarily government sponsored and subsidized and technology oriented, and the problems which are created by it. The government dominates research, and handing over much of the control of the research and development of projects to private business organizations, on which it relies for decisions. These organizations, in turn, keep their records confidential even, in some cases, from the government itself. Delicate questions thus have been placed in the hands of persons who are only very indirectly accountable to those who ultimately "foot the bill."

Government-sponsored research suffers from inefficiency and disorganization, and money is sometimes appropriated to two separate organizations studying the same problem. No government agency is charged with overall study and planning of this research, and thus it has grown like Topsy. Most research is applied and practical dealing with missiles and weaponry instead of pure science and civilian-oriented ideas. It is dominated by NASA and DOD. Certain other countries have made faster progress on civilian problems, and those in the social sciences, and have provided greater corporate subsidy for research. Furthermore, research funds are not divided equitably on a geographic basis; the midwest has been shortchanged, even though it produces most of the Ph.D.'s in the country. Concentration of research funds in a few industries and universities makes the already large and powreful even more so, and it has not benefited the smaller and weaker. The concluding chapter summarizes the changes which should be carried out to improve the situation and bring it under congressional and citizen control. The book is concluded with 20 pages of references, though many citations and examples are not footnoted.

It is hard to know how accurate a picture is painted here, since much of this information is confidential and hard to find, but most of it is probably accurate. The generalizations are often sweeping and sometimes opinionated, and the book is such to be criticized by those on the inside, though it will probably have little effect on government policy. Nevertheless, it is a useful compilation of facts documenting a picture already well-known, and it should prove useful in college and public libraries to those interested in the world of research.

JOHN F. HARVEY, Dean
Graduate School of Library Science
Drexel Institute of Technology

Can language processing be practical?
The answer is, "Yes, definitely yes."*

AUTOMATED LANGUAGE PROCESSING
The State of the Art

Edited by HAROLD BORKO, *Associate Head, Language Processing and Retrieval Staff, Research and Technology Division, System Development Corporation*

A thorough, up-to-date study of research in the use of computers to process natural languages for information purposes. Storage and retrieval, stylistic analysis, machine translation, question answering, and typesetting are covered fully, demonstrating the advances made in automated techniques being applied today in this important new area of information science.

The volume, comprised of eleven chapters written by recognized experts in the field, is divided into four main sections. The first examines the various functions basic to language processing and relates these activities to computer processing systems. The second deals with the statistical techniques of language analysis as applied to indexing and classifying documents and extracting and abstracting their contents. Also discussed is the value of statistical techniques to analyze an author's style of writing in resolving issues of disputed authorship.

The third section covers techniques and applications of syntactical analysis including the various syntactic theories and computer techniques for translating one language into another. The fourth, a single chapter by the editor, describes how computers were put to use in the typesetting and indexing of the book, providing for the reader a unique demonstration of the practicality of automated language processing.

* *From the editor's preface, in which he adds: "Unless we can develop more efficient means of communicating—sharing ideas from person to person and place to place—human progress will be inhibited."*

Contents

Language Data Processing
- Introduction (*H. Borko*)
- Language and the Computer (*L. Schultz*)
- Mathematical Models of Language (*H. P. Edmundson*)

Statistical Analysis
- Indexing and Classification (*H. Borko*)
- Extracting and Abstracting (*R. E. Wyllys*)
- Stylistic Analysis (*S. Sedelow and W. Sedelow*)

Syntactic Analysis
- Analyzing English Syntax (*D. G. Dobrow*)
- Answering Questions (*R. F. Simmons*)
- Translating Languages (*E. D. Pendergraft*)
- Designing Artificial Languages for Information Storage and Retrieval (*C. H. Kellogg*)

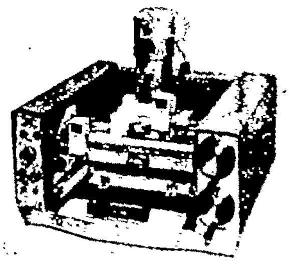
An Experiment in Language Processing, Book Indexing and Typesetting by Computer (*H. Borko*)

Conclusions
- The State of the Art

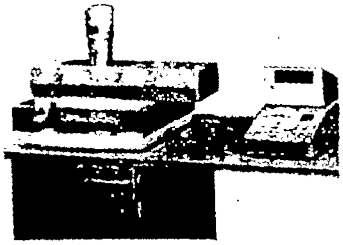
1967. Approx. 480 pages. Prob. $12.95.

Order from your bookseller or

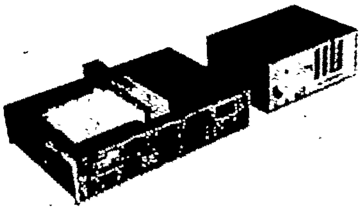
JOHN WILEY & SONS, Inc.
605 Third Avenue New York, N.Y. 10016



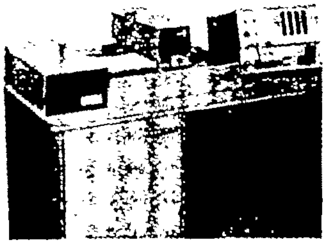
JONKER 301 data input with automatic drill action and verification of input.



JONKER 400 fully automatic data input programmed by punched cards or keyboard.

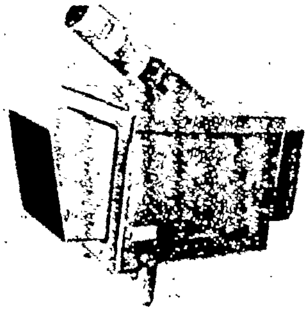


JONKER 500 fully automated scanner for statistical purposes.

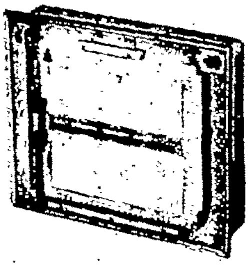


JONKER 552 fully automatic scanner which records the search results in punched cards.

JONKER 1300 reduces the regular Termatrex cards photographically to Minimatrex cards and produces an unlimited number of copies.



JONKER 1000, the reader of the Minimatrex System, the miniaturized Termatrex System, having a capacity of up to a million items.



JONKER 52 visual card reader.

Until relatively a few years ago the only effective way of automating information retrieval involved the use of the general purpose computer. Since then the JONKER Corporation has introduced a line of special purpose information search equipment, based on optical coincidence.

A recent survey of information installations using concept coordination, made under the auspices of the Information Systems Committee of the American Institute of Chemical Engineers, revealed that 40% OF THESE INSTALLATIONS NOW USE OPTICAL COINCIDENCE SYSTEMS, WHILE 31% USE COMPUTERS. FIFTEEN PERCENT USE PUNCHED CARD SYSTEMS. ONLY 12% STILL USE MANUAL SYSTEMS.

Over 600 JONKER systems are now in operation throughout the United States and as far away as Japan, South Africa and Australia. JONKER systems comprise self-contained installations as well as systems linked to punched card installations and computers. One system handles as many as 1,000 entries a day. Another serves a central library and nine satellite libraries.

Various publications of indexes based on JONKER systems, now have well over 500 users in the U.S.A. and overseas.

JONKER

Contact . . . **JONKER CORPORATION**

Main Office: Gaithersburg, Maryland (Greater Washington, D. C. area)

Telephone: 301/948-9440

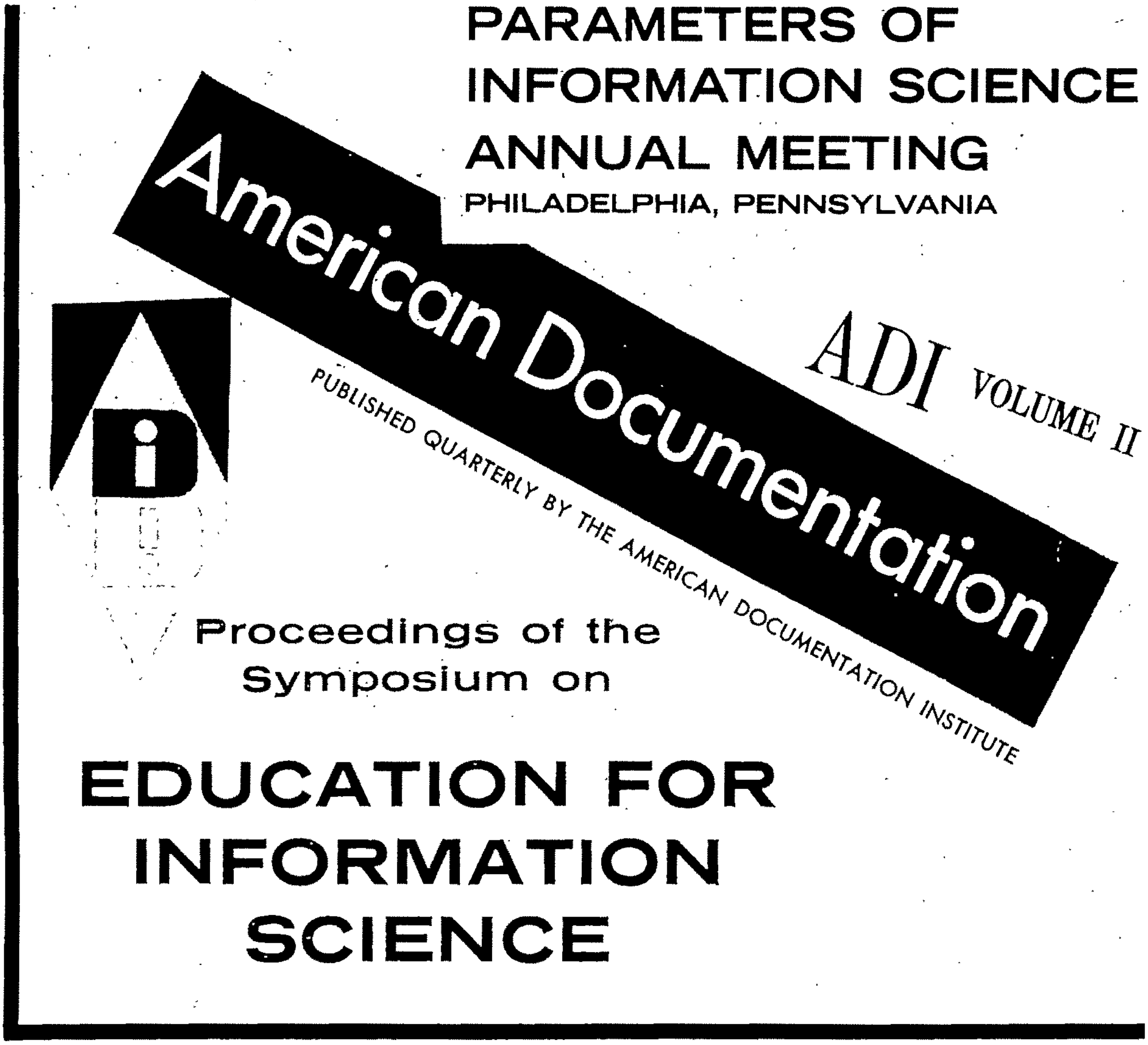
****

FIELD OFFICES: Washington, D. C. • Philadelphia, Pennsylvania • Wilmington, Delaware
Killingworth, Connecticut • Dayton, Ohio • Chicago, Illinois • Los Angeles, California

merican Documentation

APRIL
1967
Vol. 18, No. 2

PUBLISHED QUARTERLY BY THE AMERICAN DOCUMENTATION INSTITUTE



PARAMETERS OF
INFORMATION SCIENCE
ANNUAL MEETING
PHILADELPHIA, PENNSYLVANIA

American Documentation

PUBLISHED QUARTERLY BY THE AMERICAN DOCUMENTATION INSTITUTE

ADI VOLUME II

Proceedings of the
Symposium on

EDUCATION FOR
INFORMATION
SCIENCE

AMERICAN DOCUMENTATION

INSTRUCTIONS TO AUTHORS

American Documentation is a publication of the American Documentation Institute. It is a scholarly journal in the various fields in documentation and serves as a forum for discussion and experimentation. Papers already published or in press elsewhere are not acceptable. For each proposed contribution, one original and two copies (in English only) should be mailed to Mr. Arthur W. Elias, Editor, *American Documentation*, Institute for Scientific Information, 325 Chestnut St., Philadelphia, Pennsylvania 19106. The manuscript should be mailed *flat* in a suitable-sized envelope. Graphic materials should be submitted with suitable cardboard backing.

TYPES OF MANUSCRIPTS: Three types of contributions are considered for publication: full-length articles, brief communications of 1,000 words or less, and letters to the editor. Letters and brief communications can generally be published sooner than full-length manuscripts. Books, monographs, and reports are accepted for critical review. Two copies should be addressed to the Review Editor, Dr. T. Hines, 54 North Drive, East Brunswick, New Jersey.

PROCESSING: Acknowledgment will be made of receipt of all manuscripts. *American Documentation* employs a reviewing procedure in which all mansucripts are sent to two referees for comment. When both referees have replied, copies of their comments are sent to authors with the Editor's decision as to acceptability. The refereeing procedure requires about 30 days. Authors receive galley proofs with a five-day allowance for corrections. Standard proofreading marks should be employed. Reprint order forms are forwarded with galleys.

FORMAT: All contributions should be typewritten on white bond paper on one side only, leaving about 1.25 inches (or 3 cm) of space around all margins of standard, letter-size (8.5 × 11 inch) paper. Double spacing must be used throughout, including the title page, tables, legends, and references. The first page of the manuscript should carry both the first and last names of all authors, the institutions or organizations with which the authors are affiliated, and notation as to which author should receive the galleys for proofreading. All succeeding pages should carry the last name of the first author in the upper right-hand corner (0.5 inch from the top) and the number of the page.

STYLE: In general, style should follow the forms given in the Style Manual for Biological Journals (SMBJ), published for the Conference of Biological Editors by the American Institute of Biological Sciences (1964).

TITLE: The title should be as brief, specific, and descriptive as possible. Vague and unrevealing titles may delay publication.

ABSTRACT: An informative abstract of 200 words or less must be included, typed with double spacing on a separate sheet. This abstract should present the scope of the work, methods, results, and conclusions.

ACKNOWLEDGMENTS: Financial support may be listed as a footnote to the title. Credit for materials and technical assistance or advice may be cited in a section headed "Acknowledgments," which should appear at the end of the text. General use of footnotes in the text should be avoided.

GRAPHIC MATERIALS: *American Documentation* requires finished artwork. Follow the style in current issues for layout and type faces in tables and figures. A table or figure should be constructed so as to be completely intelligible without further reference to the text. Lengthy tabulations of essentially similar data should be avoided.

Figures should be lettered in black India ink. Charts drawn in India ink should be so executed throughout, with no typewritten material included. Letters and numbers appearing in figures should be distinct and large enough so that no character will be less than 2 mm high after reduction. A line 0.4 mm wide reproduces satisfactorily when reduced by one-half. Graphs, charts, and photographs should be given consecutive figure numbers as they will appear in the text; however, figure numbers and legends should not appear as part of the figure, but should be typed double spaced on a separate sheet of paper. Each figure should be marked *lightly* on the back with the figure number, author's name, complete address, and shortened title of the paper.

For figures, the originals with two clearly legible reproductions (to be sent to referees) should accompany the manuscript. In the case of photographs, three glossy prints are required, preferably 8 × 10 inches.

ORGANIZATION: In general, papers should state the background and purpose of the study, followed by details of methods, materials, procedures, and equipment. Findings, discussion, and conclusions should appear in that order. Appendixes may be employed where appropriate for extensive lists, statistics, and other supporting data.

BIBLIOGRAPHY: Accuracy and adequacy of the references are the responsibility of the author. Therefore, literature cited should be checked carefully with the original publications. References to personal letters, abstracts of verbal reports, and other unedited material may be included. If an as-yet-unpublished paper would be helpful in the evaluation of a manuscript, it is advisable to make a copy of it available to the Editor. When a manuscript is one of a series of papers, the preceding member of the series should be included in literature cited.

CITATION FORMAT:

Order: Literature cited should be sequentially numbered as cited.

Authors: Give all authors with arrangement as follows:
Elias, A. W., B. H. Weil, and I. D. Welt

Titles: Give full titles of articles in English, indicating language of original as: (In Ger.)

Journals: Journal titles should be given in full.

MONOGRAPH AND SERIAL DATA: Should be presented in order as follows: Volume, issue number, pagination, and year. The issue number should be given in parentheses if journal pagination is not continuous from issue to issue. Pagination should be inclusive. Year of publication should be given in parentheses. An example is given below:

Bishop, D., A. L. Milner, and F. W. Roper, Publication Patterns of Scientific Serials, American Documentation, 16 (No. 2): 113–21 (1965).

American Documentation is published in January, April, July, and October. One copy is included in the individual membership fee ($20.00 per year), three copies in the contributing membership fee ($100.00 per year), and up to five copies in the sustaining membership fee ($500.00 per year). Nonmembers may subscribe at $18.50 per year, postpaid in the U.S. Single copies may be purchased for $4.65 each. Communications concerning memberships, subscriptions, reprints, renewals, back issues, advertising, and changes of address should be sent to the American Documentation Institute, 2000 P Street, NW, Washington, D. C. 20036.

American Documentation is indexed in *Library Literature, Current Contents of Space, Electronic & Physical Sciences, Library Science Abstracts, Science Citation Index, Chemical Abstracts,* and *Documentation Abstracts.*

American Documentation is entered for second class mailing at Baltimore, Maryland.

American Documentation

PUBLISHED QUARTERLY BY THE AMERICAN DOCUMENTATION INSTITUTE

Vol. 18, No. 2 APRIL 1967

Published quarterly by AMERICAN DOCUMENTATION INSTITUTE and printed by Port City Press, Inc., 1323 Greenwood Road, Baltimore, Maryland 21208. Second class postage paid at Baltimore, Maryland.

Editorial

The ADI Publication Program

The cover of this issue of *American Documentation* is intended to indicate to a small degree, the proliferation of publications and projects of the American Documentation Institute since 1964. As chairman of the Publications Committee during that time, I have seen publications increase to include the Annual Review, an abstract journal, Documentation Abstracts, annual proceedings, special symposia, the memorial to H. P. Luhn, now in press and many others now being prepared or designed.

All of this activity can be ascribed to the foresight and support of four separate administrations of the Institute. It owes much to the encouragement of the Executive Director, Mr. James Bryan, and to the dedicated members of the Publications Committee, John Markus, Joe Kuney, Charles Bourne and Mary Stevens. Finally, and ultimately it is based on the contributions of the membership who write the articles and support the publications.

In 1967, the Council awarded the Editor of *American Documentation* the first honorarium ever afforded the occupant of this office. In recognition of the special responsibilities that this implies and desiring to increase the scope and size of *American Documentation* to meet the needs of the membership and of the documentation community, your Editor has declined reappointment to the Publications Committee. Another, still to be named, will take on the guidance of this growing area. I know that the loyalty and cooperation which I have received over the years will be afforded to the new incumbent and that the Publications Committee will continue the work which has had such a promising beginning.

A. W. Elias

A Film System for the Duplication of Termatrex Cards*

The Termatrex information retrieval system marketed by the Jonkers Business Machines, Inc., uses large plastic cards containing drilled holes. Duplicating the plastic cards to make identical copies for distribution requires multiple drilling and is costly, time-consuming, and error-prone. The Eastman Kodak Company has developed a method of photographing the plastic cards on cut sheet film approximately one-half the dimensions of the Termatrex cards. The film sheets can be manipulated in the same manner as the original plastic cards. The problems of supplying duplicate decks is easily handled by photographic copies of one original Termatrex deck.

C. W. BAKER, C. R. HAEFELE, and W. A. RECKHOW

Research Laboratories
Eastman Kodak Company
Rochester, New York

The Termatrex Information Retrieval System (*1, 2*) marketed by Jonker Business Machines, Inc., Gaithersburg, Md., utilizes plastic cards containing drilled holes. Each card represents a descriptor, and each hole in a given card represents a reference. Superimposition of the cards makes it possible to note which holes are in common. The locations of the holes yield reference numbers for the documents that contain descriptors in common.

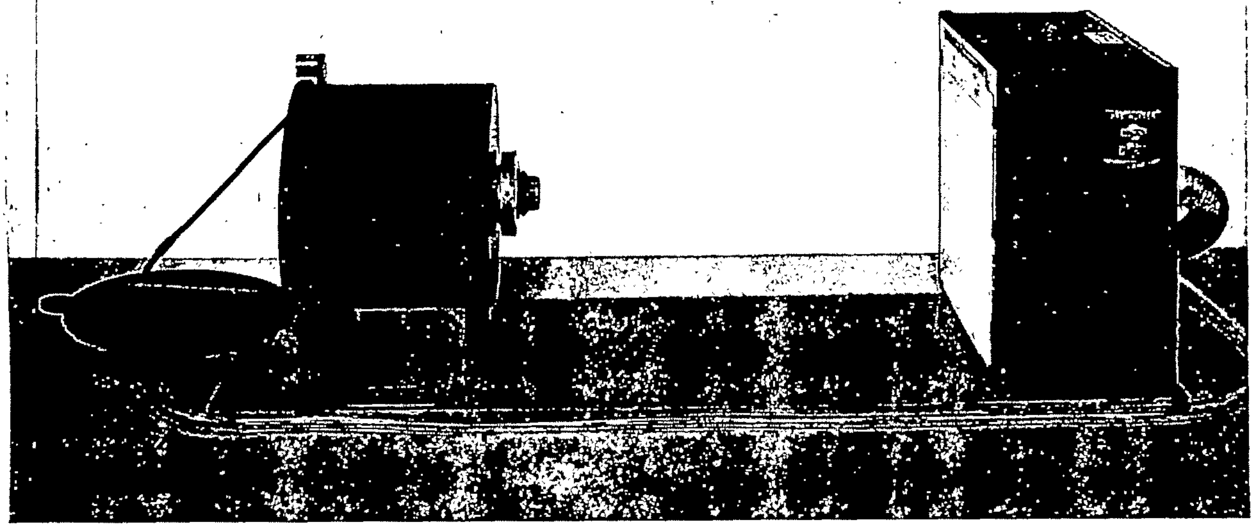


FIG. 1. Termatrex card camera (*left*) and card being photographed on the easel (*right*)

* This article is presented for information only, and nothing herein is to be construed as a recommendation or inducement to infringe patents or copyrights of others.

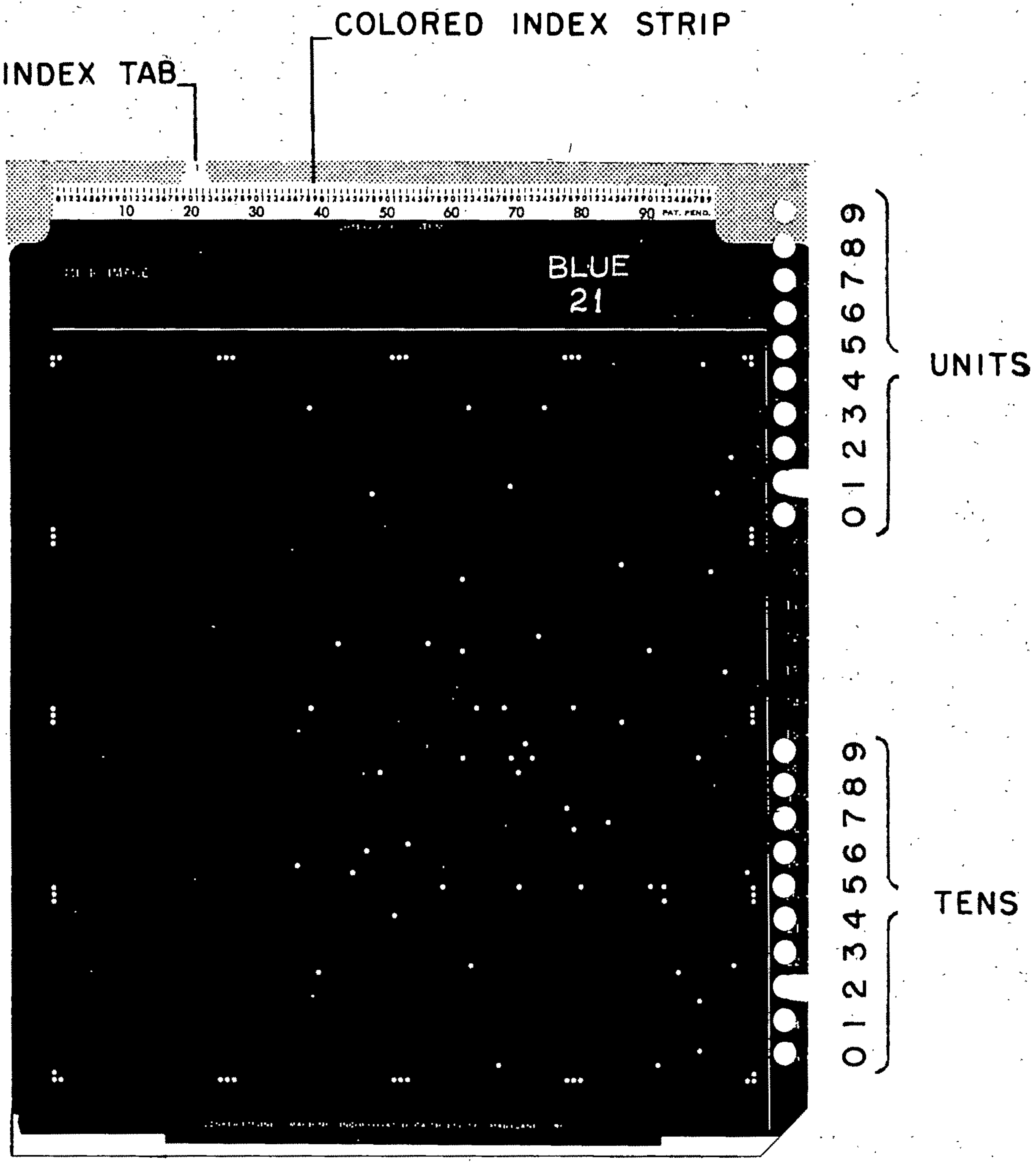


Fig. 2. Sample filmcard, showing corner cut at lower right

In the system we are using in the Research Laboratories, the top of each card is printed in one of 10 colors and has numbered positions 00–99 from left to right (Fig. 2). These numbers are located visually by tabs extending from the top of the card. The individual card is located by the color and number of the tab. One card can record reference numbers of 10,000 documents, and one deck of cards (one color) can accommodate 100 descriptors. The 10 color decks can, therefore, accommodate 1,000 descriptors.

Although the Termatrex System is versatile and economical, it has several disadvantages when duplicate decks are required. Unless all duplicates are drilled at one time, file maintenance is time-consuming, costly, and prone to error. A serious disadvantage occurs whenever decks must be updated. Either drilling equipment must be provided at each Termatrex file location, or the Termatrex cards must be shipped to a single location for drilling. One solution would be to maintain two decks, and to use one while the other is being updated. This doubles the cost and the possibility of drilling errors.

A system was designed by the Eastman Kodak Company to produce black-and-white film duplicates (filmcards) of Termatrex card decks. Filmcards can be superimposed and used in the same manner as Termatrex cards. The size of the filmcard is approximately one-half the dimensions of the Termatrex card. Drilled Termatrex cards have holes (transparent areas) and non-hole areas (opaque areas). Filmcards (film duplicates of Termatrex cards) also have holes which are transparent and non-hole areas which are opaque. Equipment was built to photograph, store, and manipulate filmcards (Figs. 1, 2, 3, 4, and 5).

The ease of handling the finished filmcards was evaluated. No registration problems were encountered, and the readability of the hole coincidences was almost as good as that of the original Termatrex cards. Use of the filmcards involved two problems: (1) selection of the desired descriptor card from the entire deck, and (2) reading of the locants or reference numbers. The color and number were printed in large letters at the top of each Termatrex card (Fig. 2). The identification was easily read on the black-and-white filmcard. All filmcards for one color were grouped together in a single module. The number order within a module may be completely random. Along the edge of each filmcard are drilled 10 holes for the tens position and 10 holes for the units position (Fig. 3). These holes were selectively notched to correspond with holes drilled in the Termatrex originals. To select "Red 53," for example, the module containing all Red-color-coded filmcards is used. Two pins are inserted in the module: at 5 in the tens position, and at 3 in the units position. The module is then inverted, and only the "Red 53" filmcard will fall out (Fig. 4). The filmcards and the module have been corner-cut to ensure proper orientation.

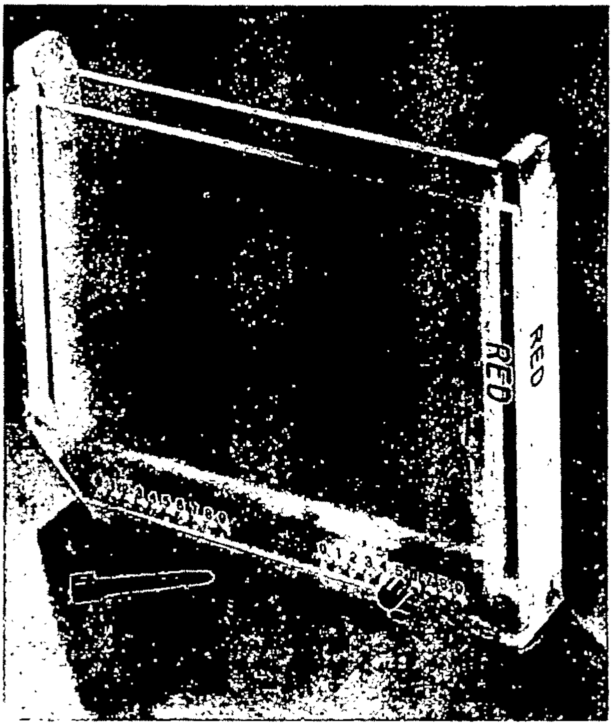


Fig. 3. Filmcard storage module

The reading of the locants for data retrieval required some means of accurately superimposing filmcards and recording the coordinates of the coincident holes. An inexpensive Plexiglas plastic viewing plaque (Fig. 5), which could be placed on an illuminator, was constructed. The filmcards selected are placed in the plaque, and over them is placed a translucent matte grid overlay. The coincident holes are marked with a soft pencil on the overlay. When the overlay is removed, the pencil marks on the grid enable the coordinates to be easily read. The marks may be erased and the grid reused. The use of replicate decks makes economical distribution of copies feasible. One central deck of Termatrex cards is kept updated, and filmcard copies are circulated to other file locations. The previous filmcard decks are then discarded. The use of filmcard decks eliminates the possibility of errors caused by multiple drilling.

References

1. Jonker, F., The New "Termatrex" Line of I.R. Systems —The "Minimatrex" Line of I.R. Systems, *American Documentation*, 14:276 (1963).

2. Sophar, G. J., Termatrex as a Tool for Storing and Searching Indexes to the Law M.U.L.L., *Newsletter, American Bar Association, Special Committee on Electronic Data Retrieval:* 1–13 (June 1964).

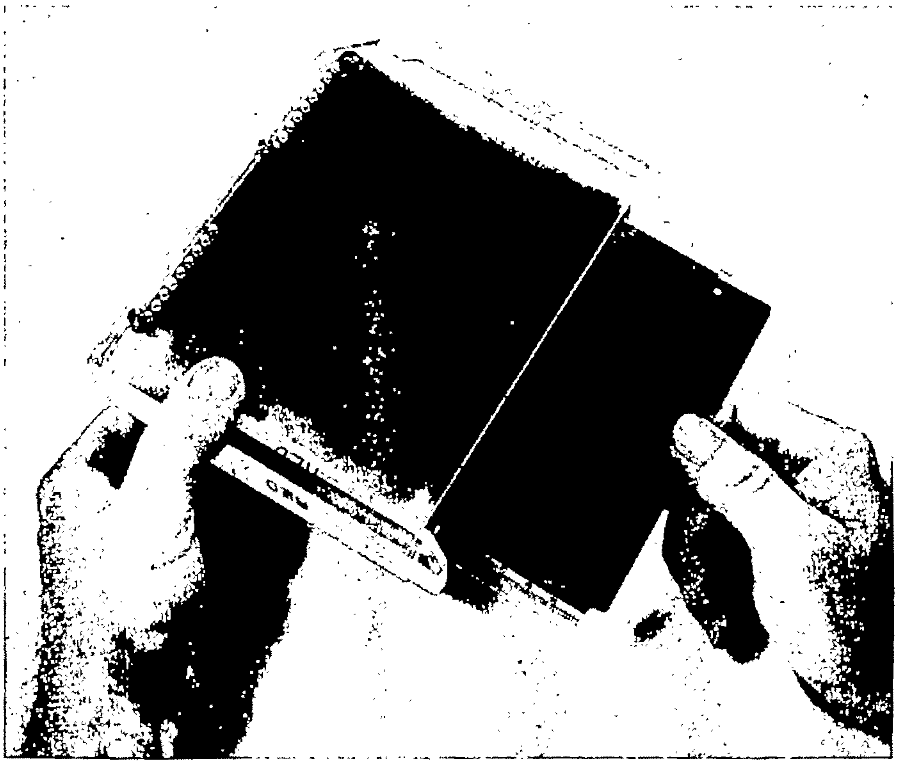
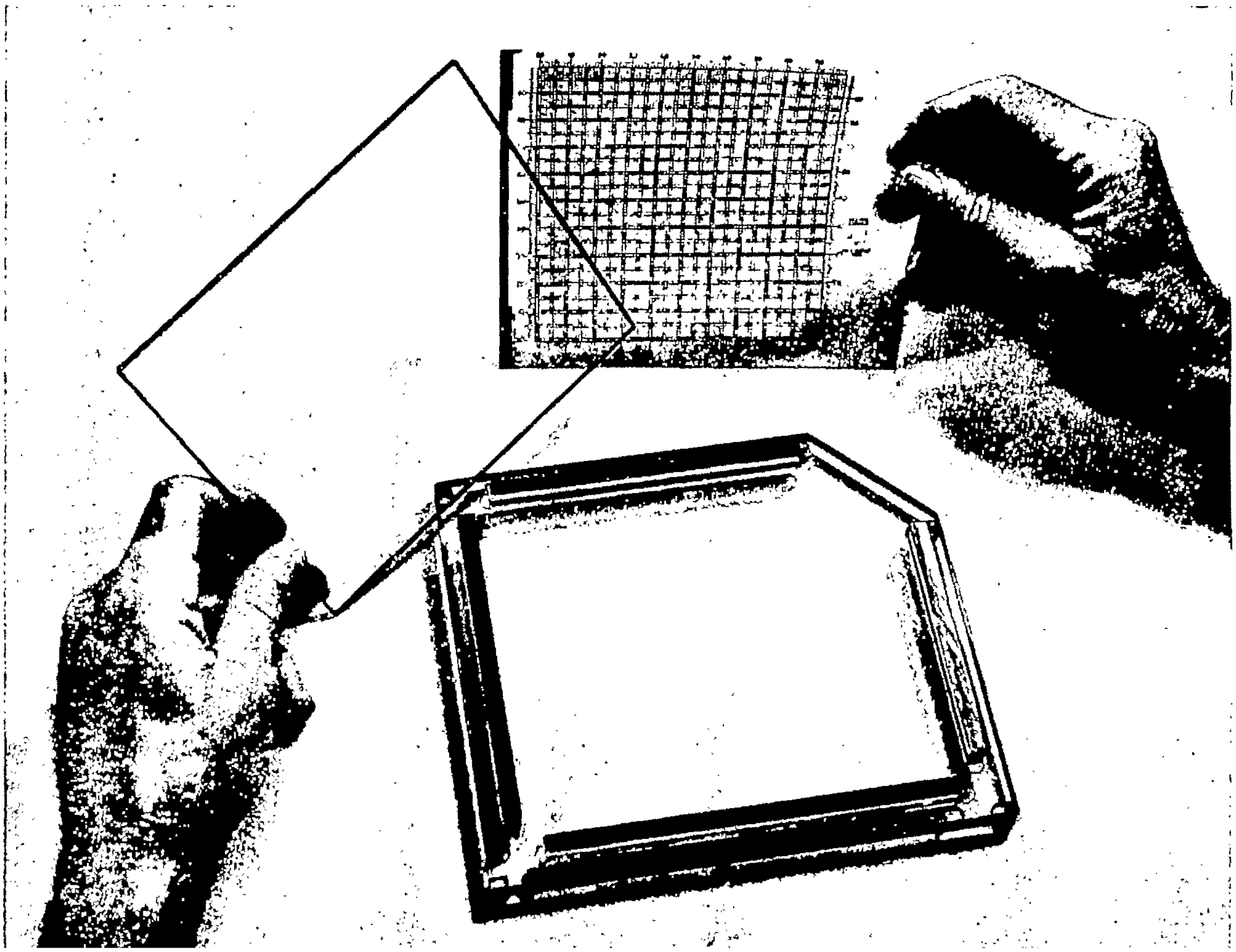


Fig. 4. Storage module showing filmcard dropout



Book-Indexes as Building Blocks for a Cumulative Index

The advantages of a cumulative subject-index, built by merging the subject-indexes of books, are illustrated; and the conditions under which such cumulation would be feasible are discussed. A mathematical model for estimating the overlap between book indexes, applied to a sample of books belonging to the same subject area, show that the overlap is surprisingly small. Some statistical properties and grammatical features of book-indexes are described in an attempt to determine how much they depart from the characteristics of book-indexes ideally suited for cumulation. Possible reasons accounting for the great variability of index-entries are discussed.

MANFRED KOCHEN and RENATA TAGLIACOZZO

Mental Health Research Institute
University of Michigan
Ann Arbor, Michigan

• Introduction

Is it feasible to cumulate the wealth of index-terms present in existing book-indexes to compile a comprehensive "catalog" for the world's book literature? This question may be of practical interest, and at the same time it may offer a good research vehicle for some insight into indexing theory. From both points of view, it is pertinent to investigate conditions under which it is profitable to merge a given collection of book-indexes to produce a cumulative index. Our first step is to identify these conditions and to observe the extent to which they prevail in current book-indexing practice.

Of the elusive process that leads from a fragment of prose to a list of words or phrases defining its semantic content, a book-index is one of the best known and most respectable products. The history of such alphabetical indexes can be traced back to the sixteenth century (*1*). Lists of terms known as "tables," "calendars," "syllabi," and "registers" were used in previous centuries to indicate the contents of books. Readers and authors of scholarly books seem to be in implicit agreement about what a book-index accomplishes, or should accomplish. Even if Lord Campbell's (1859) intention of bringing a "bill into Parliament to deprive an author who publishes a book without an index of the privilege of copyright and moreover to subject him for his offense to a pecuniary penalty" (*1*, p. 28) was never carried into action and a good number of indexless books are still published, the value of indexes is universally recognized. Yet it is impossible to find in the literature any satisfactory description of standards for differentiating good indexes from bad indexes, or for that matter, any convincing explanation of the indexing process.

A few treatises and handbooks on index-making have been published at various times (*2*) to illustrate current procedures for compiling indexes. While focusing their attention on technical details, such as alphabetization of index entries, cross-references, hierarchy of headings, they omit problems of a more fundamental nature, e.g., concerning the structure of an index. In vain does one try to obtain from these publications an answer to questions of the following kinds: (a) How does one select the terms suitable to index a particular fragment of text? or (b) What constitutes an index entry?

Current research on indexing (*3*, *4*) aims, on the whole, at systematic procedures to produce an index from a corpus of text. Investigators engaged in this kind of research start from the assumption that the products of traditional indexing practices are generally inconsistent and have high cost/effectiveness ratios. Their investigations aim to introduce revolutionary changes in indexing practice and are more relevant to texts being generated from today on than to the already-indexed past literature. In fact, none of the proposed indexing schemes have as yet been proved to yield a sufficiently low cost/effectiveness ratio to warrant converting the *texts* of *past* literature into machine-readable form for reindexing by computer. Nor has any proposed scheme been sufficiently forward looking and favorably evaluated to persuade the intellectual community to abandon traditional indexing practice in its favor (*5, 6*).

The possibility of a simple system to bridge the transition between currently used indexing practices and future, radically improved practices may thus deserve serious consideration. Poor as they may be, existing indexes do contain a wealth of information that could be used to create a vastly more effective index at a reasonable cost. Moreover, the problem solved and the procedures used in this operation could provide the base on which to build realistic and radically improved indexing systems.

• Why a Cumulated Index?

A cumulated index built by merging the subject-indexes of a large number of books would be equivalent, in many respects, to a "subject catalog." This, however, is not to be visualized as the room full of card-filled drawers now found in a library. Nor need it necessarily be in book form, like the National Union Catalog. This is too large and too costly to produce, update, and distribute as widely as desirable. Instead, consideration should be given to the effective use of time-sharing technology, both in the cumulation as well as in the use of the "end-product." Indeed, the "end-product" would be a *continually* changing cumulation of book-indexes stored in large-capacity, digital-computer stores. Users would interrogate this catalog from remote access stations, which would be connected to such a digital store for on-demand interrogation. For example, a user seeking an operational definition of the unit of time would have to know that "frequency standard" or "standard second" are the closest "jargon" terms to use; he is not likely to find much of relevance if he performs the search by using subject-headings like "Time," "Clocks," etc. If he does somehow hit on "frequency standard," he may be referred, among other references, to the American Institute of Physics Handbook, sec. 5, p. 112, where he would immediately learn that the standard second is 1/86,400 of a mean solar day; that a quartz-crystal oscillator at the Bureau of Standards is used to monitor a 5-cycle radio pulse at a frequency of 1 kc at the beginning of the last second of each minute, to an accuracy of 1 microsecond. If this is what he was really after, rather than, say, an explanation of the Heisenberg Uncertainty Principle, or the smallest time interval that the latest nuclear counters can record, he would be led to this as rapidly as he could use the console of the time-sharing computer, i.e., within minutes. Of course, if the user of the cumulated index is to obtain an immediate and relevant response to his query, then at least one of the books that contain the statement he needs should have an appropriate index entry. (Ideally each of the books should have an appropriate index entry.)

What are the advantages of using a cumulation of book indexes over current practices? In current practice, to locate a relevant passage on, say, "frequency standards," a user would first have to select a subject-heading under which a book containing this passage is likely to be cataloged. There is quite a variety of subject headings the user may think of and an even greater variety he would not think of. Supposing that the appropriate subject headings were found, the user might then have to scan the titles or even the tables of contents of so many irrelevant books that he might well doubt if the value of meeting his need is worth all this effort. This sense of doubt is, of course, very much a question of attitude. Convenience of using the index by suitable display techniques, timing of responses to queries, etc. can affect this attitude as much as the quality of indexing and may often spell the difference between the user's bothering to go after a piece of information or judging its value to be less than the effort to hunt for it.

If the advantages of a cumulated book-index were limited to convenience of use and speed of retrieval, it is doubtful that from a practical point of view attempts to construct such an index could be justified. The superiority of a cumulated index, however, is based primarily on the fact that, through it, a large number of points of access to the content of books would become available. In existing book catalogs the subject headings assigned to each book are very few (in the Library of Congress Catalog, for instance, there are about 1.6 per book). In the Cumulated Index there could be as many as there are index entries in the book-index, i.e., a few hundred. This means that when seeking information on a specific issue, one would not be forced to search under a *generic* subject-heading, which may or may not give access to the desired information (and which would certainly produce a large number of irrelevant answers), but would have direct access to the information via the *specific* index-entry.

• Some Quantitative Aspects of Cumulation

Consider a collection of t books. A fraction p of these contain passages relevant to an index term x, e.g., $x=$ "frequency standard." Pick one of the pt relevant books at random and let r denote the conditional probability that it contains the term x in its index given that it is relevant. This relevant book may fail to contain x because the author did not index the passage, or because he assigned a term other than x to it, etc. If the querist is satisfied with any one of the pt relevant books, he will fail to be satisfied if none of the relevant books have x in their index. If the events of several relevant books each failing to contain x in their index can be assumed to be statistically independent, then the probability that each of the pt relevant books does not contain x in its index is $(1-r)^{pt}$. The conditional probability of retrieving a book from the merged indexes of all t books, given that the book is relevant, is $1-(1-r)^{pt}$. This is also the hit-rate (7) or recall-ratio (5).

Now pick one of the $t-pt$ irrelevant books at random. Let r' be the probability that it does not contain x in its index. It could, erroneously, contain x in the index if the author, for example, indexed a passage under x even though the passage merely mentions x or something about x, yet giving no substantive information. Or it could contain x as an appropriate index term; but if x means something different to the indexer and to the querist, the book may be irrelevant. If the assumption that such errors in several books occur independently is valid, then the probability that all the $t-rt$ irrelevant books fail to contain x is $r'^{(t-pt)}$; and the probability that at least one of them contains x is $1-r'^{t-pt}$. With the help of Bayes' rule we can now express the conditional probability of a book being relevant given that it is among those that contain x in its index in the merged index of t books (also known as acceptance-rate (7) or relevance-ratio (5)) as

$$P(\text{relevant/retrieved}) =$$

$$\frac{P(\text{retr./relev.})\ P(\text{relev.})}{P(\text{retr./relev.})\ P(\text{relev.}) + P(\text{retr./irrelev.})\ P(\text{irrelev.})}$$

$$=\frac{[1-(1-r)^{pt}]p}{[1-(1-r)^{pt}]p+[1-r'^{t-pt}](1-p)}$$

$$=\frac{1}{1+\frac{1-\mathrm{p}}{\mathrm{p}}[1-r'^{t-pt}][1-(1-r)^{pt}]^{-1}}$$

This quantity increases to p as t increases; that is, if the indexes to many books are merged into a combined index, then the chances of finding a relevant passage by looking under x are not much better than when selecting one of the t books at random. It can, however, be kept quite large when t is not too big.

Compare these results with the corresponding quantities under the present system. Given the index term x, it is necessary to pick first a suitable subject heading y. Then as many books cataloged under y must be retrieved and scanned for the presence of x in the book-index as necessary to obtain a match. Consider, as before, a collection of t books, pt of which contain a needed passage. Let r and r' be defined as before. Now let R be the probability that a book will be retrieved given that it contains x in its index, and R' the probability that a book will not be retrieved given that it does not contain x in its index. These two quantities depend upon the "thesaurus" in which the relation between an index-term like x="frequency standard" and various subject-headings is specified. They also depend upon the cataloger's judgment in assigning subject-headings to books.

It can be shown that the probability that a relevant book is retrieved is now $Rr+(1-R')(1-r)$. This does not depend on t, as might be expected. To derive this, it was assumed that the conditional probability of a book being relevant, given that it is retrieved through the catalog and that x is in its index, is equal to the conditional probability of a book being relevant given that x is in its index. In a similar manner, it can be shown that the probability of an irrelevant book being retrieved is now $R(1-r')+(1-R')r'$.

To take a sample calculation, let $t=100$ books on nearly the same subject. We assume that by chance alone 1 out of such a hundred contains the needed passage, so that $p=.01$. Suppose further that $R=R'=.9$ and $r=r'=.95$. Then the hit-rate and acceptance-rate for the present system are 0.86 and 0.14, respectively. For a merged index they would be 0.95 and 0.60, respectively.

To estimate the size of the merged index, which in turn determines the time needed to look up an index term, let n_1 be the average number of index entries in a book. Let n_2 be the average number of index entries common to any two books. Generally, let n_k be the average number of index entries, each of which appears in each of a random sample of k books. The average number of new terms added by adjoining a second index to the starting index will be n_1-n_2. The total number of different entries in the merger of the two will be $n_1+(n_1-n_2)$. The number of *new* terms contributed by adjoining a third index will be $n_1-2n_2+n_3$; by a fourth, $n_1-3n_2+3n_3-n_4$. Generally, the number of *new* entries contributed by the $(k+1)^{st}$ index is

$$\binom{k}{0}n_1-\binom{k}{1}n_2+\binom{k}{2}n_3-\ldots\pm\binom{k}{k}n_{k+1}$$

or

$$\sum_{j=0}^{k}(-1)^j\binom{k}{j}n_{j+1}\text{ for }k=0,1,2,\ldots,t$$

The total number N_t of *different* index entries in the cumulated index resulting from merging t book-indexes is

$$N_t=\sum_{k=0}^{t-1}\left(\sum_{j=0}^{k}(-1)^j\binom{k}{j}n_{j+1}\right)$$

If it could be assumed that $n_j=As^j$ where A is a constant greater than 1 and s a constant such that $0<s<1$, then it is easily shown with the help of the binomial theorem and the expression for the sum of a geometric series that

$$N_t=\frac{A[1-(1-s)^t]}{s}$$

As t increases, N_t converges exponentially from below to A as a horizontal asymptote. Thus, A is the ultimate size the index could attain under this assumption. A would be smaller if all books were on the same topic than if they were not. The smaller s, the more rapidly N_t converges to A; the larger s, the more slowly. Thus $1/s$ indicates the degree to which different authors use the same index terms. If, even in a narrow field, authors varied greatly in the index terms they assigned—even to the same passage—then s would be close to 1; if they adhered to a standard set of index terms, s might be close to 0. The latter condition would more likely hold in physical sciences where the jargon is standardized; the former in historical studies, where there is less agreement about how to label a finding.

The assumption, $n_j = As^j$, is not tenable however. To determine how n_k, the number of index entries which appears in each of a k randomly selected books, decreases with k, we selected 10 books in the field of "learning." Of the 45 possible pairs, we selected 20 at random and counted the number of index terms-in-common for both members of each pair (Table 1). To compensate for the variance in the size of the various indexes, we divided the number of terms-in-common by the product of the number of index terms in each of the two indexes. This has been shown to be an unbiased, minimum-variance estimate. The median value of these 20 numbers was 0.543×10^{-4}. The median rather than the mean had to be used because the distribution is quite skew.

Of the $\binom{10}{3} = 120$ possible book triples, 30 were selected at random. One triple had 10 index entries in common, another triple had 7, another 5, two triples had 4, five triples had 3, four had 2, ten triples had 1, and six triples had no index entries in common. Dividing the number of entries-in-common by the product of the sizes of the three indexes, the median of these 30 numbers was 0.039×10^{-6}.

Repeating this procedure for quadruples gave a median of 0.00525×10^{-8}. In the case of quintuples, only 1 of 30 had one index entry common to all 5.

The average number of index entries/book was about 400. We therefore estimated n_2 by $(400)^2 \times 0.543 \times 10^{-4} = 8.68$, n_3 by $(400)^3 \times .038 \times 10^{-6} = 2.43$, and n_4 by $(400)^4 \times 0.00525 \times 10^{-8} = 1.34$. On a log-log plot, log n_j vs. log$_j$ is very nearly linear for these points, and a good fit is provided by

$$n_j = 38.6 j^{-2.5} \text{ for } j = 2, 3, 4, \ldots ; \; n_1 = 400$$

Without computing

$$N_t = \sum_{k=0}^{t-1} (400 - k(38.6) \cdot 2^{-2.5} + \frac{k(k-1)}{2} (38.6) \cdot 3^{-2.5}$$
$$- \ldots \pm 38.6(k+1)^{-2.5})$$

it is easy to see that the index resulting from a cumulation of t books will have fewer than t times the average number of entries in one book index. This number will, however, grow significantly with t and will not reach an asymptote as t gets large, as was the case for the assumption $n_j = As^j$.

The following observations are noteworthy.

1. The extent to which several indexes contain precisely the same index entries is (surprisingly) small.
2. The rate of decrease in the number of index terms-in-common to k indexes with k is less than exponential; it is such that for large k the number is still significantly high.

Condition 1 does not exclude the feasibility of cumulating book-indexes. It tells us merely that the cumulated index will be quite large and will continue to grow as we add indexes. If we wish to maintain a limiting size, we will have to purge it continually of less significant entries.

To make a large cumulated index usable, the index terms will have to be richly interconnected, cross-referenced through relations of synonymy, part-whole, generic-specific, etc. Moreover, our estimates were based on requiring that a pair of index terms be matched in every (linguistic) detail to be called common to two indexes. The effect of less rigid criteria for the similarity or relatedness of two index terms on the production and use of the cumulated index is being investigated separately.

TABLE 1. Similarities among subject-indexes of 10 books on "learning"

Number of identical index-entries	Sample from all possible combinations of the 10 book-indexes			
	Pairs ($N = 20$)	Triplets ($N = 30$)	Quadruplets ($N = 30$)	Quintuplets ($N = 30$)
30	1			
19	1			
17	1			
15	1			
12	1			
11	3			
10	2	1		
9	1			
8	1			
7	3	1		
6				
5		1		
4	3	2		
3	1	5	1	
2	1	4	1	
1		10	7	1
0		6	21	29

• Conditions Favorable to Cumulation of Book Indexes

What are the properties of a book-index which would make it suitable for merging with other book-indexes to produce a comprehensive search tool for the total book literature? We have already pointed out that, in order to give a relevant response to a query, the cumulative index should have an appropriate index entry to the book which contains the relevant information. Thus, our first requirement for a suitable book index is:

> *Rule 1.* Each passage in the book likely to answer an anticipated query should be referenced in the index.

Conformity to this rule is as difficult to test as it is important. The book indexer, frequently the author of the book, hastily and casually assigning "index terms" while proofreading the final galley, usually cannot anticipate all the queries his book may answer. Perhaps he can anticipate a few hundred of the queries he intends to answer. This may account for half of all the queries that would be asked if really good access to all the potential answers were available. Perhaps the other half of all such queries could not have been anticipated by anyone at the time the book was written. Perhaps people other than the author can more easily think of possible queries. Nonetheless, most existing books provide some references to the most important passages in the book, and this could serve as a useful starting point.

A second important prerequisite for cumulating book-indexes is the following:

> *Rule 2.* Each index entry should be in a standard linguistic form.

The first problem encountered when one compares two indexes, in fact, consists of determining whether two linguistically different index terms are the same. For example, "frequency standards" and "standards of frequency" are probably equivalent. Linguistic preferences of this type vary considerably in book indexes. Yet there may be certain statistical regularities in the use of various grammatical forms. The double noun form as in "frequency standards" may, for example, occur much more frequently than the equivalent noun-preposition-noun form. This might suggest a canonical form into which to transform all index-entries for purposes of comparison.

The third desirable property of book-indexes can be defined as follows:

> *Rule 3.* If passages from different books treat similar topics with the same level of specificity, index-terms having the *same* level of specificity should be assigned to those passages.

Deviations from this rule are probably the most troublesome and the least amenable to standardization (*8*). When selecting an index entry appropriate to a passage, different indexers exhibit great freedom in the choice of specificity level. Some of them may use very generic, all inclusive, one-noun categories, while others may prefer to index a similar passage with a very specific entry, in which the same noun appears amidst a large number of modifiers. Under these conditions, estimating the number of index-entries common to two indexes becomes very difficult. Does the difference in specificity reflect a difference in text specificity, or a difference in indexers' preference? To what degree does the size of the indexes determine the level of specificity of the index-entries? It is clear that if two books of approximately the same size have indexes of different length, the larger index will have a tendency to include more specific entries than will the shorter index. In which way, then, will level of specificity, size of index, and size of book be related? It is not difficult to see that from the interplay of the various factors outlined here a large variability of index entries, even within the same subject field, is to be expected.

• Some Properties of Book-Indexes

Of a sample of 66 books selected at random from a research library collection in the biological and social sciences, 10 books (15%) had no subject index at all. The size of the subject index was distributed over the remaining 56 books as shown in Table 2.

From this table we can see that about 55% (or over one-half) of the books have a subject index of less than 400 entries. An index of this size, which takes at most five pages of two-column print, can be considered a small index. The medium-size indexes (those including from 400 to 800 entries) contribute 27% (or a little over one-fourth) to the total, and finally only 7% of the sample consists of larger indexes (between 800 and 1,200 entries). The remaining 11% of the indexes spread widely between 1,200 and 7,000 entries.

The ratio between number of entries in the index and number of pages in the book (density of index entries) varies, in our sample, between 0.15 and 4.71. The distribution is shown in Table 3.

We see, then, that over 90% of the books in the sample have an index density lower than 3.00. The average density is between 1.00–1.50 index entries for each page of book. Considering the variety of factors that influence the ratio between index size and book size (e.g., the restrictions imposed by publishing policies, the differences in level of redundancy of the text and in degree of specificity of the index lexicon), the variance of the index density does not seem extremely high.

If one disregards the "function" words (for the most part prepositions), the majority of index entries are composed of either one, two, or three words. We found

TABLE 2. Distribution of index size in a sample of 56 scientific books from various fields

Index size (Number of entries* per index)	Number of books	Percent of sample
00– 200	14	55
200– 400	17	
400– 600	10	27
600– 800	5	
800–1000	2	7
1000–1200	2	
1200–1400	1	
1400–1600	0	
1600–1800	1	
1800–2000	1	
2000–2200	0	
2200–2400	1	
2400–2600	0	
2600–2800	1	
....	...	
6800–7000	1	

* We define an "entry" to be a word or a group of words followed by one or more numbers specifying the book location (page) containing the passage referred to.

TABLE 3. Distribution of index entries density in a sample of 56 books

Density (entries/page)	Number of books
0– .50	5
.50–1.00	16
1.00–1.50	12
1.50–2.00	9
2.00–2.50	7
2.50–3.00	3
3.00–3.50	1
3.50–4.00	1
4.00–4.50	1
4.50–5.00	1

that in a sample of 10 indexes of small and medium size, the total one-, two-, and three-term entries account, on the average, for over 90% of the subject index. The distribution of the percentages is given in Table 4.

In the top five books of the table, which have less than 240 entries, the percentages of the one-term entries are greater than the corresponding percentages for the five bottom books having more than 240 terms per index. This makes sense if one thinks that when available space is limited, priority is given to one-term entries, which usually are on a higher level of generalization than are the two- and three-term entries. (Compare, for instance, the entries "index," "book index," and "alphabetical book index".) Only larger indexes, in which space is no problem, can afford to include a large percentage of three-term index entries.

The inverse relation between size and percentage of the one-term entries, which tends to keep the absolute number of one-term entries within narrow numerical limits, seems interesting. Possibly, the pool of technical terms from which single-term index entries are drawn is, for each scientific area, rather limited in size. This would explain why the total number of one-term index entries is not much higher in the larger indexes than in the smaller ones. The multiple-term index entries then would be formed by combining these characteristic technical terms with those derived from the much larger pool of terms belonging to the everyday (i.e., nonscientific, nontechnical) language.

TABLE 4. Percent index entries of different length in a sample of 10 books

	Total index entries No.	One-term entries %	Two-term entries %	Three-term entries %	Residue %
Book A	150	54.6	38.7	6.7	..
Book B	150	32.7	46.6	18.7	6.0
Book C	151	48.4	27.8	13.9	9.9
Book D	235	42.1	44.5	11.9	1.2
Book E	240	68.3	27.5	4.2	..
Book F	369	15.7	47.7	29.3	7.3
Book G	401	34.9	53.6	10.0	1.5
Book H	522	33.7	44.2	19.2	2.9
Book I	646	13.6	41.5	28.3	16.6
Book J	671	17.6	37.8	24.3	20.3

We were interested in finding out which grammatical forms are prevalent in index entries. On this subject, manuals for index compilation are explicit, even if brief: nouns or noun phrases should be used as index entries. We found that most of the indexers abide by these rules. This is not enough, however, to keep the various indexes within the limits of a standard grammatical format, as anybody can realize at first glance by comparing a small number of indexes.

In our sample of 10 book-indexes, we identified 25 types of grammatical combinations which account, on the average, for about 98% of the one-, two-, and three-term entries, and for about 92% of the total number of entries.

These 25 categories were differentiated on a purely morphological basis, without giving consideration to the fact that some of them are semantically equivalent and are used interchangeably in current language (e.g., noun-noun vs. noun-preposition-noun).

The extent to which the 25 grammatical forms exhaust all the possible combinations of terms appearing in index-entries depends, to a large degree, on the size of the index. Small indexes, which contain a limited number of three-term entries and very few four-term or longer entries, fit almost completely within our 25 categories; while large indexes, which have a broader repertoire of multiple-term combinations, show a larger residue of unclassified entries.

The percentages of the various grammatical forms occurring in each of the 10 indexes of our sample were calculated. The 10 higher-ranking grammatical forms are shown in Table 5; and their average percentage, over the 10 indexes of the sample, is given.

Altogether the combinations shown in Table 5 account for over 80% of the total number of entries and can

TABLE 5. The 10 higher-ranking grammatical forms occurring in a sample of 10 book-indexes

Grammatical form of index entry	Average percent of total number of entries
1. Noun	34.51
2. Adjective-noun	16.13
3. Noun-noun	12.80
4. Noun-preposition-noun	6.40
5. Adjective-noun-noun	3.28
6. Adjective-noun-preposition-noun	2.44
7. Noun-preposition-noun-noun	1.96
8. Gerund	1.69
9. Noun-noun-noun	1.57
10. Present participle–noun	1.35
	82.13

therefore be considered to represent the main body of a book-index of small or medium size. As we have pointed out, large indexes present a more varied array of grammatical forms; but, on the other hand, they occur less frequently, so that they do not alter considerably the rank-order of grammatical forms presented in Table 5.

• Conclusions

Our results show that subject-indexes of different books have very little in common. This is quite surprising, especially for our sample of 10 books that not only belong to the same subject field and have the same orientation, but are also close in date of publication. This latter fact is of some significance in favoring the assumption that these 10 books revolve around the same problems and concepts and that they were indexed on the basis of the same current practices. How can one explain the striking differences exhibited by their indexes?

If we want to attempt to answer this question, we should go back to examine the properties of the average book-index, and see if we can identify the various factors that may account for their variability.

If a subject-index were merely an alphabetical list of key-words, books belonging to the same field and treating the same subject would probably share a considerable number of identical index-entries. Key-words, in fact, are derived primarily from the pool of technical words characteristic of a particular field, which is rather restricted in size and includes primarily standardized words. It is true that a certain number of index-entries are single nouns, and are therefore to be considered equivalent to key-words. But one-word entries, although accounting for a considerable portion of the total number of entries in small indexes, are only a minor part of medium or large indexes. In large indexes, the percentage of one-word entries goes down to 5% or less. When we compare book-indexes of different sizes, then, the matches generated by coincidence of one-term entries are limited in number.

Most index-entries are formed by a noun together with one or more modifiers. Although we have not attempted to find out what the upper limit in the number of modifiers is, we know that often index-entries are phrases of considerable length. Changes in number, relative position, and grammatical form of the modifiers provide a large array of combinations and permutations of index-entries. We have seen that, even if we limit the analysis to entries of between one and three terms and omit variations in relative positions of the terms, we end up with a list of 25 combinations, which exhaust only 92% of the possible variations offered by our sample. The variability of index-entries is further augmented by the addition of generic nouns that have practically no retrieving power but are used to indicate relationships between two or more terms of the entry (e.g., importance, cause, effect, factor, similarity).

We have already pointed out in the section titled "Conditions Favorable to Cumulation of Book Indexes" that an important determinant of index-entries variability is the different level of specificity used by different indexers to index similar text passages. In some cases this difference in specificity level is required by the particular situation. A short index, for instance, may have room only for generic entries. On the other hand, *specific* entries may be appropriate for a small index, given the fact that the subject field of the book is very restricted and only *specific* problems or concepts are discussed in the text. We are not in a position, at this stage, to define the ideal level of specificity for a book-index. All we can do is to point out the problems created by large discrepancies in level of specificity of index-entries. Consider, for instance, the low probability of finding an exact match, either in another book-index, or in a query, for the following index-entries: (a) "preexperimental habits and difficulty in assessing commonality of behavioral laws;" or (b) "operational definitions as basis for taxonomy of learning." (These are two nonatypical index-entries in a recent volume publishing the proceedings of a symposium on the psychology of human learning.)

Another crucial factor responsible for variability of index-entries is the selection from the text of those "significant passages" that deserve to be indexed. This depends so much on the competence of the indexer, on his personal judgment, his attitude, and the way he conceives the process of indexing, that it is not surprising if the results are at times hardly comparable with those produced by another indexer. We can refer to recent studies on reliability of indexing which show a rather low agreement among subjects (and in the same subject at different times) in selecting "representative" sentences from scientific articles (*9, 10*).

We must remember that differences in methods of indexing are not the only determinants of book-indexes variability. Obviously, differences inherent in the books themselves (i.e., difference in text) are responsible for differences in the indexes. There are many ways in which books may differ. If we confine our analysis to scientific books, we can detect three main types of text differences responsible for variations in subject-indexes. One is the orientation of the book (by which we mean primarily the kind of readers for whom the book is written); the second is the level of specificity of the text; and the third is, of course, the subject of the book. The first variable can perhaps be equated with the degree of technicality of the text. A book of zoology written for grammar school students is certainly less technical than a textbook of zoology for college students. On the other hand a textbook of zoology may be as *technical* as a book on protozoa, but probably is less *specific* (at least on the topic of protozoa).

The subject of books is the main determinant, naturally, of index variability. Only in books from the same or partially overlapping fields can we expect to find similarity of index-entries.

We can summarize the preceding discussion by saying that we can detect six factors affecting book-indexes variability, three of them related to index production, the other three to text production:

Related to index production
Size and grammatical form of entries
Level of specificity of entries
Selection of text-passages to be indexed

Related to text production
Level of specificity of text
Orientation of book
Subject of book

As an answer to the question raised in the introduction, a cumulation of book-indexes, although feasible, may not be so simple and practical as it might have been if the overlap among book-indexes had been found to be greater. Our findings suggest further study into the relationships among index entries in order to estimate the limiting efficiency with which a cumulation of book indexes could be produced and used.

References

1. Wheatley, H. B., *What is an Index?* 2nd ed., Longmans, Green and Company, London, 1879.
2. Collison, R., *Indexing Books,* T. De Graft, New York, 1962.
3. Artandi, S., Automatic Book Indexing by Computer, *American Documentation,* 15:250–257 (1964).
4. O'Connor, J., Mechanized Indexing Methods and Their Testing, *J. Assoc. Comp. Mach.,* 11:437–449 (1964).
5. Cleverdon, C. W., and J. Mills, The Testing of Index Language Devices, *ASLIB Proceedings,* 15:106–130 (1963).
6. Lancaster, F. W., and J. Mills, Testing Indexes and Index Language Devices: The ASLIB Cranfield Project, *American Documentation,* 15:4–13 (1964).
7. Kochen, M., Toward Document Retrieval Theory: Relevance Recall Ratio for Text Containing One Specified Query Term, *in* H. P. Luhn (Ed.), *Automation and Scientific Communication,* 3:439–442 (1963).
8. Loukopoulos, L., Indexing Problems and Some of Their Solutions," *American Documentation,* 17:17–25 (1966).
9. Rath, G. J., A. Resnick, and T. R. Savage, The Formation of Abstracts by the Selection of Sentences. Part I. Sentence Selection by Men and Machines, *American Documentation,* 12:139–141 (1961).
10. Resnick, A., The Formation of Abstracts by the Selection of Sentences. Part II. The Reliability of People in Selecting Sentences, *American Documentation,* 12: 141–143 (1961).

Computer-Produced Microfilm Library Catalog

The philosophy, production, and cost-effectiveness of a computer-generated library catalog is described. This catalog is unique in that it utilizes direct computer to microfilm composition techniques, employing the Stromberg Carlson 4020. Cost, user acceptance, and by-product capabilities are stressed.

W. A. KOZUMPLIK and R. T. LANGE

Technical Information Center
Lockheed Missiles & Space Company
Sunnyvale, California

Production of the library catalog depends on the correct blend of the following ingredients: capability, cost, and user acceptance. This axiom was tested in the nineteenth Century when the card catalog displaced catalogs in book form.

With the advent of computer technology, the book catalog has come full circle within a century. To be sure, the first catalogs produced by computers were in card form, which was a natural evolutionary advance when passing from manual to automated systems. Irrespective of format, the point to be stressed is that each of these computer-produced catalog systems, whether card or book, was required to stand the test of capability, cost, and user acceptance. And it is precisely this test that the third catalog system passed convincingly.

This third system is the computer-produced library catalog in microfilm form. It is a product of off-the-shelf hardware and of programming excellence. Administrators of large specialized libraries as well as directors of research will be particularly interested in the microfilm system because it is the least costly and the most advanced, effectively user-oriented catalog system in operation today.[1]

Such a system was installed in the Technical Information Center (TIC) of Lockheed Missiles & Space Company (LMSC) and became fully operational in July 1966. The capabilities of this new system far outshine its predecessors and will be discussed in detail here. Besides achieving these advanced capabilities, the present system actually returns a moderate savings in comparison to its forerunners. Having passed the capability and costs tests, the system was presented to the public—the operational as well as the administrative user, that is, the scientist on the one hand and the librarian on the other—with favorable results. The user found that look-up time was greatly reduced and that the system was easy to operate.

Lockheed's experience has affirmed user acceptance to be broadly based and has brought requests to TIC management to install a microfilm catalog in R&D oriented buildings. This can be done at no extra computer costs and is beneficial whenever high-priced scientists and engineers are located at a considerable distance from the library. Obviously convenience of look-up on the premises of his own building will revolutionize the researcher's work and should improve the quality of his product and prevent unwanted duplication; this, after all, is the *raison d'être* of the special library. The point to be emphasized is that a catalog that is located a few feet from the researcher's work station dynamically improves his accessibility to the company's cataloged literature resources.[2]

The computerized catalog system installed and operating at LMSC delivers the following products in accordance with design requirements:

1. An updated library catalog in microfilm form
2. A listing of new publications added, using the keyword-in-title (KWIT) format
3. An updated report of open-entry items contained in the library
4. A source authority list, with appropriate cross-references

[1] On-line real time dialog library catalog systems, admittedly more powerful, are still excessively costly.

[2] The increased volume of use and reuse of cataloged information, of course, inevitably introduces the problem of logistics; that is, the need to acquire or generate multiple copies to satisfy simultaneous multiple user needs.

5. A subject authority list, with "see" and "see also" references appropriate

These products are processed quarterly with the exception of the KWIT, which is issued semi-monthly. Basic to all products is the system-derived and magnetic-tape stored master file on all publications contained in the library together with the ability to delete, add, or change records on the master file as determined by the controlling organization, namely, the library management.

It is worthy of note that these products listed can be delivered for several separate libraries within the same computer-processing cycle. At LMSC for instance, these products are currently separately generated for two collections while two additional collections are in the process of converting to this low-cost retrieval system.

The system originates with source documents being keypunched and forwarded to the computer (Fig. 1). These records are of various types, the most predominant being catalog additions and deletions. The cards are generated onto magnetic tape and sorted. The input transactions are subjected to certain editing requirements, reformatted, and exploded into various multiple records. This explosion is based upon the number of tracings in each document. During this phase of the operation, the documents that pertain to new publications are also generated on a separate tape that produces the keyword-in-title listing.

The next step in the system operation is to sort the edited-exploded transactions into the same sequence as the master file. These sorted transactions are processed against the master file to produce the updated catalog and an updated master file. Also during this pass, a tape is generated that produces the source and the subject authority listings.

The user-related visible components of the computer-produced microfilm library catalog system are microfilm cartridges and a microfilm reader together with the semi-monthly KWIT entitled *New Reports & Books*.

The catalog's 1,051,060 look-up points or entries are organized in six sections: source, title, author(s), contract number, subjects, and report numbers/call number (Figs. 2, 3, 4, 5, 6, 7, 8). Both reports and books which heretofore had been cataloged and shelf-ordered according to separate systems, resulting in separate catalogs, are now for the first time integrated into a single catalog. The 16 mm microfilm compressing these million-plus retrieval points are loaded into 40 cartridges; each cartridge contains 100 ft of film on which are exposed 1,800 two-column pages of computerized catalog text processed by the SC 4020. Each page contains approximately 14 entries. Altogether, this is a significant compression of text and space since the million-plus records when in card form had previously occupied 720 standard library catalog drawers. The cartridges are housed in an 80-compartment rack that stands next to the reader on a 60×30 in. work table (Fig. 9). Each cartridge is labeled as to contents. The labels are colored differently to provide visual ease in distinguishing the six separate sections of the catalog.

Any on-shelf automatic microfilm reader and associated cartridges may be used to display the microfilmed text. For the installation at LMSC the Bell & Howell Microfilm Reader, Model 531, was selected because it provides not only visual reading comfort with its zoom lens and its three intensities of lamp brightness but also speed. The latter is derived from the use of the Bell & Howell patented automatic no-rewind cartridge. The user simply removes the cartridge after his look-up; the next user of that cartridge does not have to rewind the film but merely commences his search to the left or right as the case may be. The zoom lens enlarges text size up to 100%.

A complete, cumulated, and corrected microfilm catalog is produced quarterly: Computer processing time together with duplicate microfilm processing, label generation, cartridge loading; and distribution to the operating location takes ten work days. The decision to schedule production on a quarterly basis rather than bi-monthly or even monthly was founded on production costs and computer availability. For instance, the bi-monthly production cycle would cost $3,000 more annually.

Between periodic microfilm catalog production runs, users are kept informed of titles added to the collections by the semi-monthly computer product in KWIT format, *New Reports & Books*. This is a variant of Bell Telephone Laboratory's BEPIP program and is structured in two parts, Title, and Bibliography. It is not a retrospective retrieval tool of any great effectivity but serves basically to announce works newly added during the period reported to the inventory of literature resources that are available to qualified users. For the user whose approach is by subject, author, or contract number, the three-month gap in currency of the catalog denies him access to the latest information in the inventory. But when the user's approach is by source, report number, or title, the KWIT is moderately helpful; consequently, this type of user will be but mildly adversely affected in his exploitation of the most recent resources.

Queuing problems caused by multiple simultaneous catalog utilization are forestalled by installing multiple readers and catalogs at a ratio of two to one for the library's clientele and one to one for library technical services staff. The cost of these added equipments and components is more than offset by savings derived from text compression. In actuality, the savings in compression paid for the four catalogs and eight readers installed on library premises for use of scientist/engineers and for the five catalogs and five readers for use of library staff. In addition, the microfilm system operates at a net saving of $13,000 annually because (1) card filing costs are avoided; (2) there are no catalog cases to purchase; and (3) there is a 200% saving in space to house the microfilm installation.

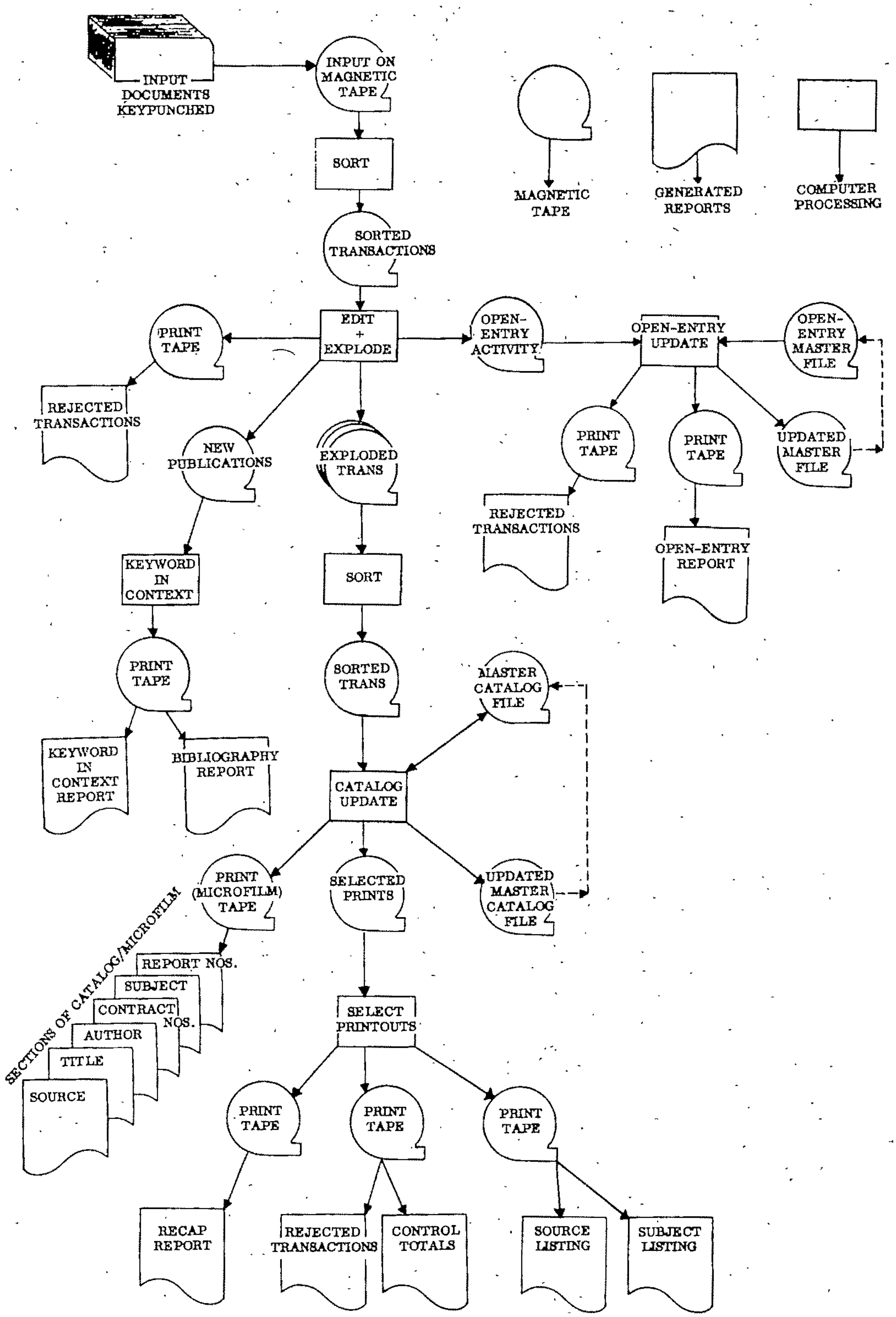
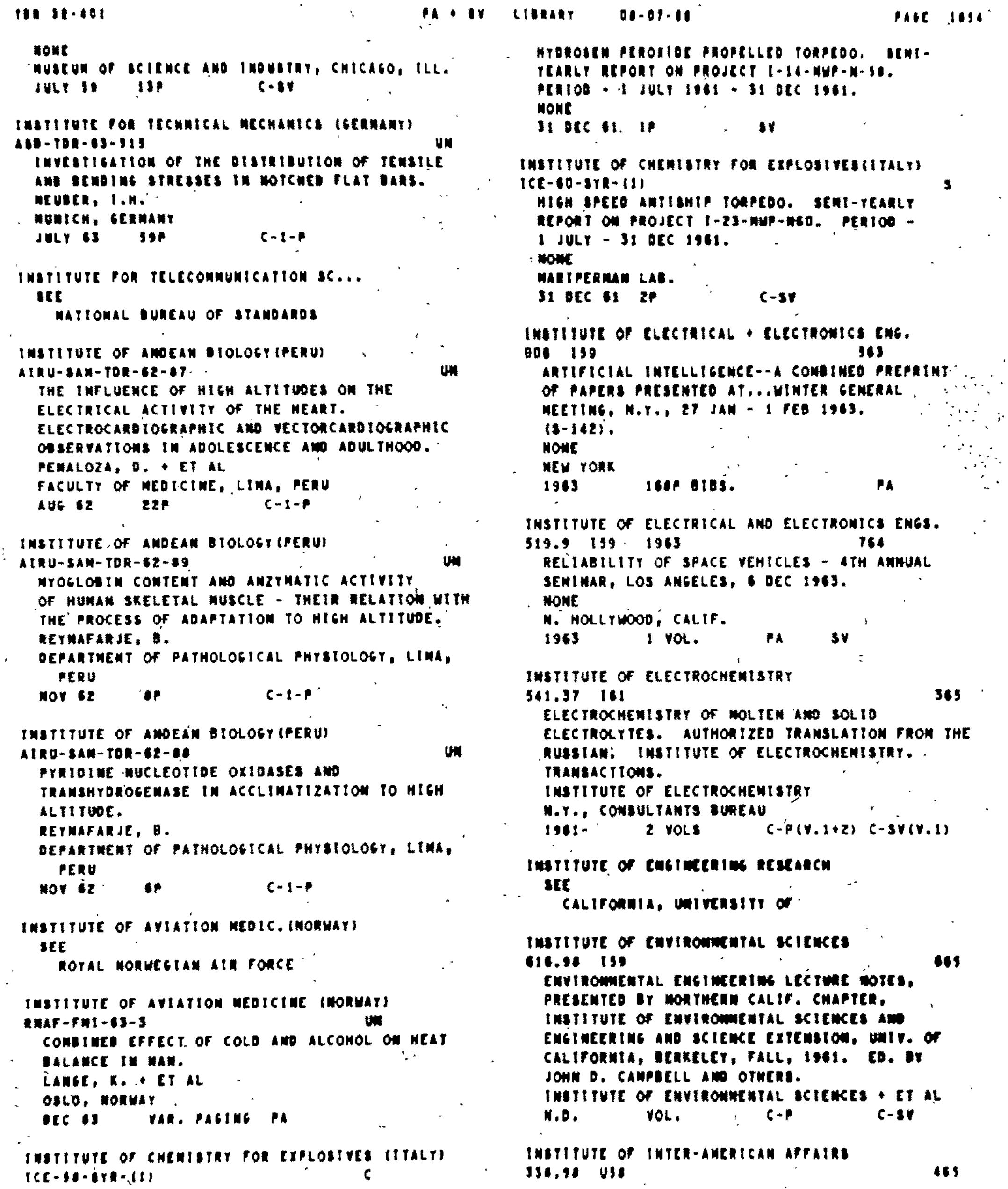
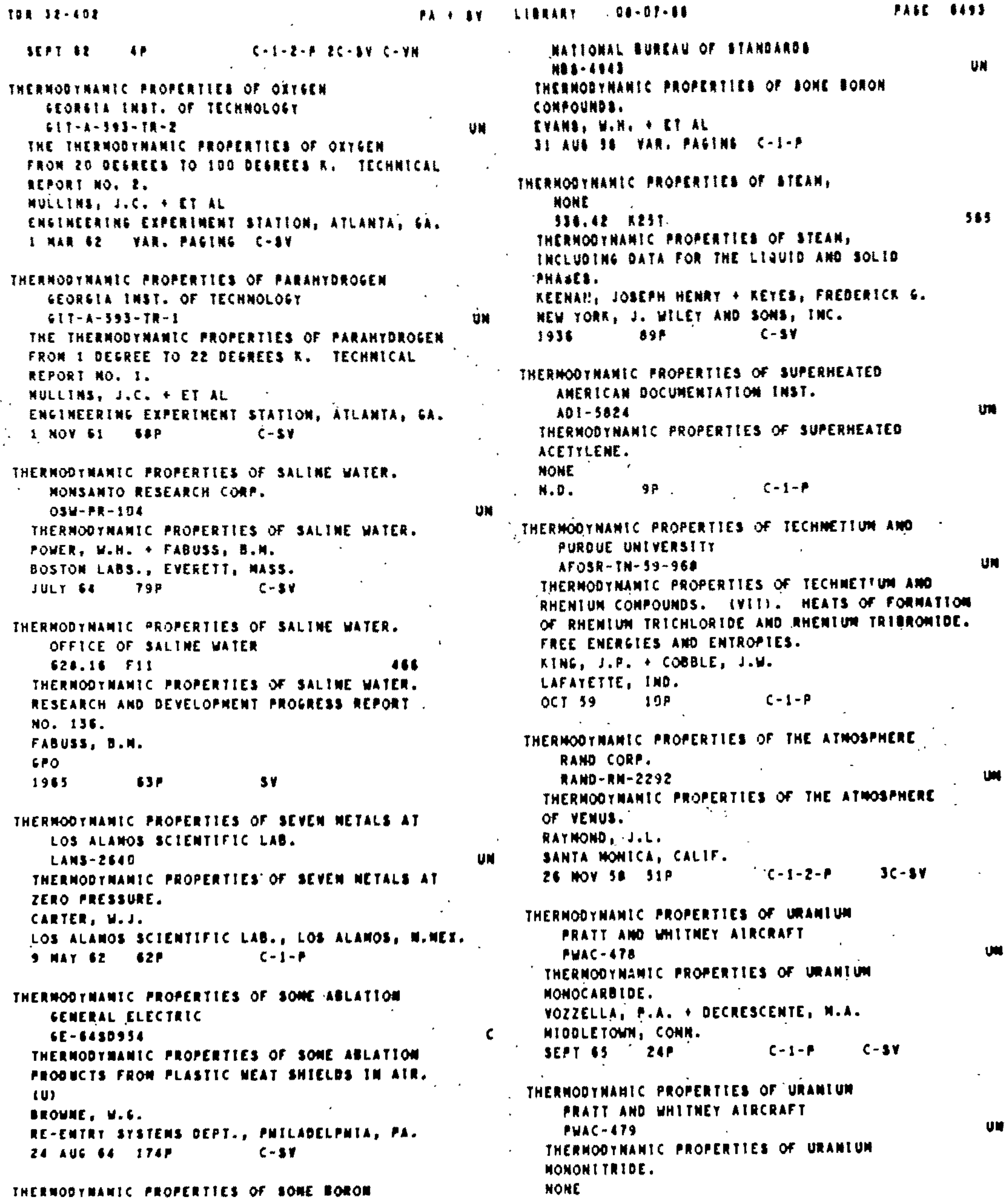


FIG. 1. LMSC Library Catalog, Data system flow



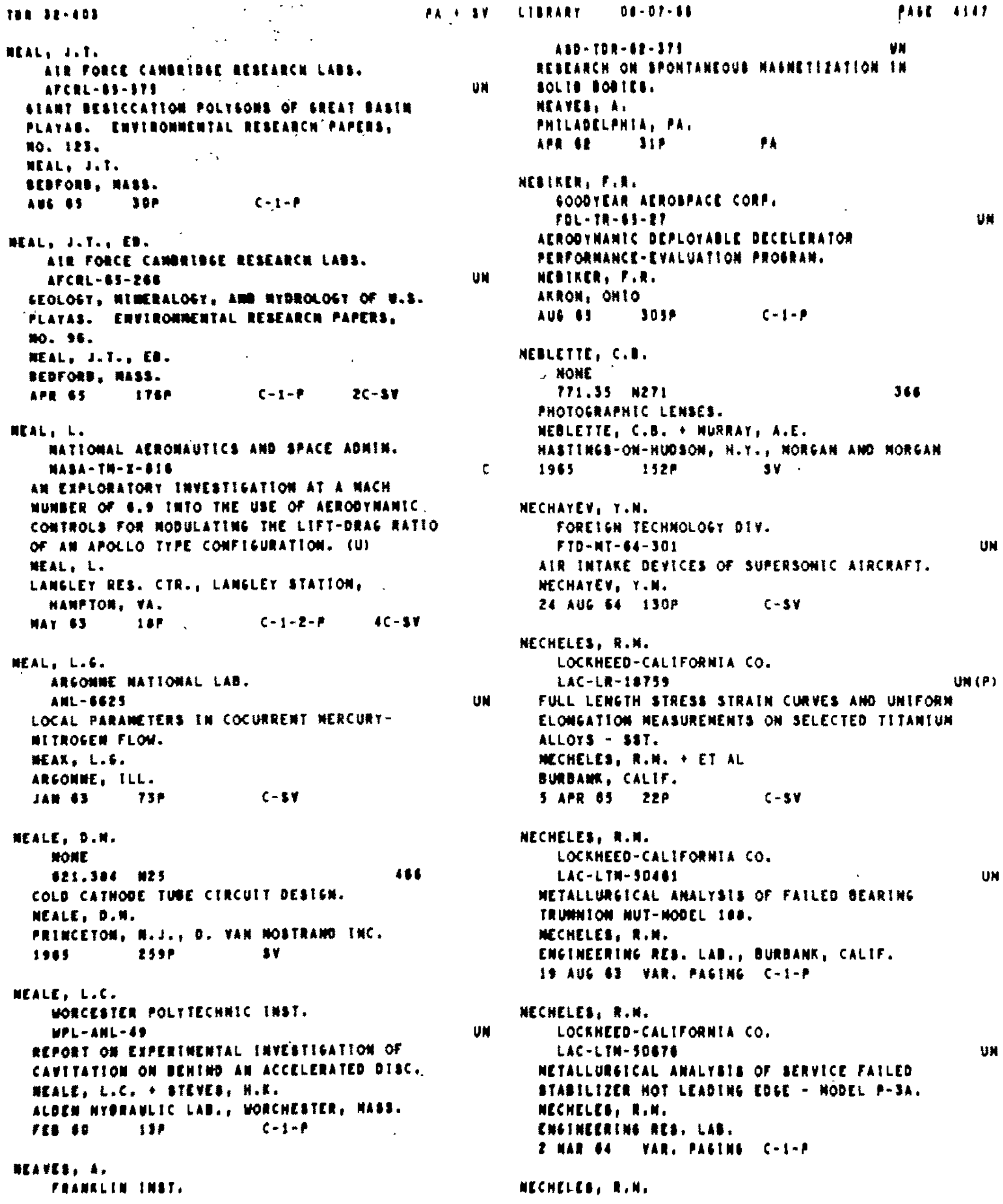
```
TDR 32-401                            PA + SV   LIBRARY     08-07-66                    PAGE 1034

  NONE                                                HYDROGEN PEROXIDE PROPELLED TORPEDO.  SEMI-
  MUSEUM OF SCIENCE AND INDUSTRY, CHICAGO, ILL.       YEARLY REPORT ON PROJECT I-14-NWP-N-59.
  JULY 59     13P           C-SV                      PERIOD - 1 JULY 1961 - 31 DEC 1961.
                                                      NONE
INSTITUTE FOR TECHNICAL MECHANICS (GERMANY)           31 DEC 61  1P          SV
ASD-TDR-63-515                             UN
  INVESTIGATION OF THE DISTRIBUTION OF TENSILE      INSTITUTE OF CHEMISTRY FOR EXPLOSIVES(ITALY)
  AND BENDING STRESSES IN NOTCHED FLAT BARS.        ICE-60-SYR-(1)                                 S
  NEUBER, I.H.                                        HIGH SPEED ANTISHIP TORPEDO.  SEMI-YEARLY
  MUNICH, GERMANY                                     REPORT ON PROJECT I-23-NWP-N60.  PERIOD -
  JULY 63     59P           C-1-P                     1 JULY - 31 DEC 1961.
                                                      NONE
INSTITUTE FOR TELECOMMUNICATION SC...                 MARIPERMAN LAB.
  SEE                                                 31 DEC 61  2P          C-SV
    NATIONAL BUREAU OF STANDARDS
                                                    INSTITUTE OF ELECTRICAL + ELECTRONICS ENG.
INSTITUTE OF ANDEAN BIOLOGY(PERU)                   806  I59                              563
AIRU-SAM-TDR-62-87                         UN         ARTIFICIAL INTELLIGENCE--A COMBINED PREPRINT
  THE INFLUENCE OF HIGH ALTITUDES ON THE              OF PAPERS PRESENTED AT...WINTER GENERAL
  ELECTRICAL ACTIVITY OF THE HEART.                   MEETING, N.Y., 27 JAN - 1 FEB 1963.
  ELECTROCARDIOGRAPHIC AND VECTORCARDIOGRAPHIC        (S-142).
  OBSERVATIONS IN ADOLESCENCE AND ADULTHOOD.          NONE
  PENALOZA, D. + ET AL                                NEW YORK
  FACULTY OF MEDICINE, LIMA, PERU                     1963       160P BIBS.               PA
  AUG 62      22P           C-1-P
                                                    INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGS.
INSTITUTE OF ANDEAN BIOLOGY(PERU)                   519.9  I59  1963                      764
AIRU-SAM-TDR-62-89                         UN         RELIABILITY OF SPACE VEHICLES - 4TH ANNUAL
  MYOGLOBIN CONTENT AND ANZYMATIC ACTIVITY            SEMINAR, LOS ANGELES, 6 DEC 1963.
  OF HUMAN SKELETAL MUSCLE - THEIR RELATION WITH      NONE
  THE PROCESS OF ADAPTATION TO HIGH ALTITUDE.         N. HOLLYWOOD, CALIF.
  REYNAFARJE, B.                                      1963       1 VOL.      PA     SV
  DEPARTMENT OF PATHOLOGICAL PHYSIOLOGY, LIMA,
    PERU                                            INSTITUTE OF ELECTROCHEMISTRY
  NOV 62      8P            C-1-P                   541.37  I61                                365
                                                      ELECTROCHEMISTRY OF MOLTEN AND SOLID
INSTITUTE OF ANDEAN BIOLOGY(PERU)                     ELECTROLYTES.  AUTHORIZED TRANSLATION FROM THE
AIRU-SAM-TDR-62-88                         UN         RUSSIAN.  INSTITUTE OF ELECTROCHEMISTRY.
  PYRIDINE NUCLEOTIDE OXIDASES AND                    TRANSACTIONS.
  TRANSHYDROGENASE IN ACCLIMATIZATION TO HIGH         INSTITUTE OF ELECTROCHEMISTRY
  ALTITUDE.                                           N.Y., CONSULTANTS BUREAU
  REYNAFARJE, B.                                      1961-      2 VOLS       C-P(V.1+2) C-SV(V.1)
  DEPARTMENT OF PATHOLOGICAL PHYSIOLOGY, LIMA,
    PERU                                            INSTITUTE OF ENGINEERING RESEARCH
  NOV 62      6P            C-1-P                     SEE
                                                        CALIFORNIA, UNIVERSITY OF
INSTITUTE OF AVIATION MEDIC.(NORWAY)
  SEE                                               INSTITUTE OF ENVIRONMENTAL SCIENCES
    ROYAL NORWEGIAN AIR FORCE                       616.98  I59                                665
                                                      ENVIRONMENTAL ENGINEERING LECTURE NOTES,
INSTITUTE OF AVIATION MEDICINE (NORWAY)               PRESENTED BY NORTHERN CALIF. CHAPTER,
RNAF-FMI-63-3                          UN             INSTITUTE OF ENVIRONMENTAL SCIENCES AND
  COMBINED EFFECT OF COLD AND ALCOHOL ON HEAT         ENGINEERING AND SCIENCE EXTENSION, UNIV. OF
  BALANCE IN MAN.                                     CALIFORNIA, BERKELEY, FALL, 1961.  ED. BY
  LANGE, K. + ET AL                                   JOHN D. CAMPBELL AND OTHERS.
  OSLO, NORWAY                                        INSTITUTE OF ENVIRONMENTAL SCIENCES + ET AL
  DEC 63      VAR. PAGING  PA                         N.D.       VOL.        C-P         C-SV

INSTITUTE OF CHEMISTRY FOR EXPLOSIVES (ITALY)       INSTITUTE OF INTER-AMERICAN AFFAIRS
ICE-59-SYR-(1)                         C            338.98  US8                                465
```

FIG. 2. Sources



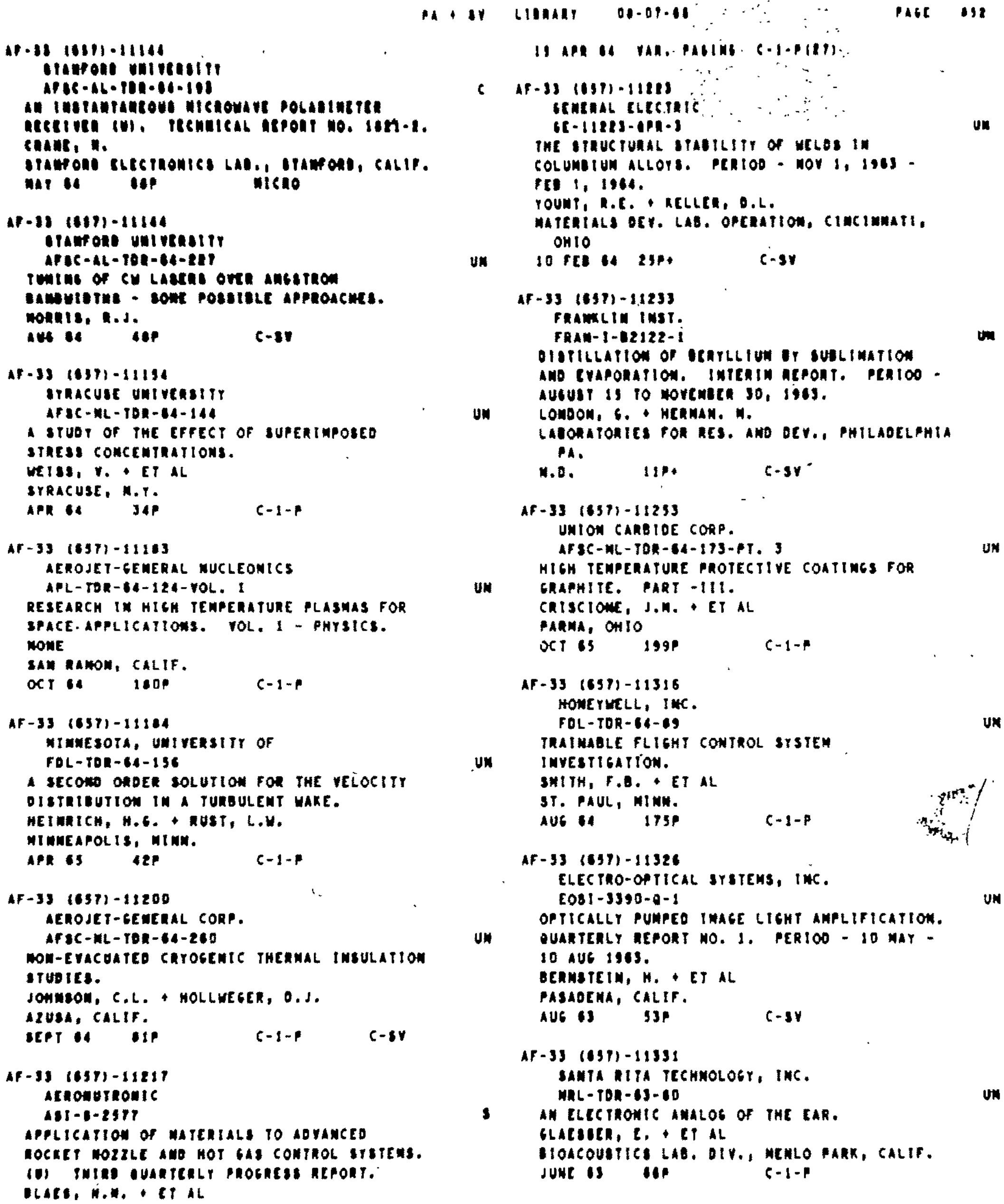
```
TDR 32-402                          PA + SV   LIBRARY   08-07-66                PAGE  0493

  SEPT 62     4P           C-1-2-P 2C-SV C-VN

THERMODYNAMIC PROPERTIES OF OXYGEN
    GEORGIA INST. OF TECHNOLOGY
    GIT-A-393-TR-2                                UN
  THE THERMODYNAMIC PROPERTIES OF OXYGEN
  FROM 20 DEGREES TO 100 DEGREES K.  TECHNICAL
  REPORT NO. 2.
  MULLINS, J.C. + ET AL
  ENGINEERING EXPERIMENT STATION, ATLANTA, GA.
  1 MAR 62    VAR. PAGING  C-SV

THERMODYNAMIC PROPERTIES OF PARAHYDROGEN
    GEORGIA INST. OF TECHNOLOGY
    GIT-A-393-TR-1                                UN
  THE THERMODYNAMIC PROPERTIES OF PARAHYDROGEN
  FROM 1 DEGREE TO 22 DEGREES K.  TECHNICAL
  REPORT NO. 1.
  MULLINS, J.C. + ET AL
  ENGINEERING EXPERIMENT STATION, ATLANTA, GA.
  1 NOV 61    68P          C-SV

THERMODYNAMIC PROPERTIES OF SALINE WATER.
    MONSANTO RESEARCH CORP.
    OSW-PR-104                                    UN
  THERMODYNAMIC PROPERTIES OF SALINE WATER.
  POWER, W.H. + FABUSS, B.M.
  BOSTON LABS., EVERETT, MASS.
  JULY 64     79P          C-SV

THERMODYNAMIC PROPERTIES OF SALINE WATER.
    OFFICE OF SALINE WATER
    628.16  F11                           466
  THERMODYNAMIC PROPERTIES OF SALINE WATER.
  RESEARCH AND DEVELOPMENT PROGRESS REPORT
  NO. 136.
  FABUSS, B.M.
  GPO
  1965        63P          SV

THERMODYNAMIC PROPERTIES OF SEVEN METALS AT
    LOS ALAMOS SCIENTIFIC LAB.
    LAMS-2640                                     UN
  THERMODYNAMIC PROPERTIES OF SEVEN METALS AT
  ZERO PRESSURE.
  CARTER, W.J.
  LOS ALAMOS SCIENTIFIC LAB., LOS ALAMOS, N.MEX.
  9 MAY 62    62P          C-1-P

THERMODYNAMIC PROPERTIES OF SOME ABLATION
    GENERAL ELECTRIC
    GE-64SD954                                    C
  THERMODYNAMIC PROPERTIES OF SOME ABLATION
  PRODUCTS FROM PLASTIC HEAT SHIELDS IN AIR.
  (U)
  BROWNE, W.G.
  RE-ENTRY SYSTEMS DEPT., PHILADELPHIA, PA.
  24 AUG 64   174P         C-SV

THERMODYNAMIC PROPERTIES OF SOME BORON
    NATIONAL BUREAU OF STANDARDS
    NBS-4943                                      UN
  THERMODYNAMIC PROPERTIES OF SOME BORON
  COMPOUNDS.
  EVANS, W.H. + ET AL
  31 AUG 56   VAR. PAGING  C-1-P

THERMODYNAMIC PROPERTIES OF STEAM,
    NONE
    536.42  K25T                                  565
  THERMODYNAMIC PROPERTIES OF STEAM,
  INCLUDING DATA FOR THE LIQUID AND SOLID
  PHASES.
  KEENAN, JOSEPH HENRY + KEYES, FREDERICK G.
  NEW YORK, J. WILEY AND SONS, INC.
  1936        89P          C-SV

THERMODYNAMIC PROPERTIES OF SUPERHEATED
    AMERICAN DOCUMENTATION INST.
    ADI-5824                                      UN
  THERMODYNAMIC PROPERTIES OF SUPERHEATED
  ACETYLENE.
  NONE
  N.D.        9P           C-1-P

THERMODYNAMIC PROPERTIES OF TECHNETIUM AND
    PURDUE UNIVERSITY
    AFOSR-TN-59-960                               UN
  THERMODYNAMIC PROPERTIES OF TECHNETIUM AND
  RHENIUM COMPOUNDS.  (VII).  HEATS OF FORMATION
  OF RHENIUM TRICHLORIDE AND RHENIUM TRIBROMIDE.
  FREE ENERGIES AND ENTROPIES.
  KING, J.P. + COBBLE, J.W.
  LAFAYETTE, IND.
  OCT 59      10P          C-1-P

THERMOOYNAMIC PROPERTIES OF THE ATMOSPHERE
    RAND CORP.
    RAND-RM-2292                                  UN
  THERMOOYNAMIC PROPERTIES OF THE ATMOSPHERE
  OF VENUS.
  RAYMOND, J.L.
  SANTA MONICA, CALIF.
  26 NOV 58   51P          C-1-2-P     3C-SV

THERMODYNAMIC PROPERTIES OF URANIUM
    PRATT AND WHITNEY AIRCRAFT
    PWAC-478                                      UN
  THERMODYNAMIC PROPERTIES OF URANIUM
  MONOCARBIDE.
  VOZZELLA, P.A. + DECRESCENTE, M.A.
  MIDDLETOWN, CONN.
  SEPT 65     24P          C-1-P      C-SV

THERMODYNAMIC PROPERTIES OF URANIUM
    PRATT AND WHITNEY AIRCRAFT
    PWAC-479                                      UN
  THERMODYNAMIC PROPERTIES OF URANIUM
  MONONITRIDE.
  NONE
```

FIG. 3. Titles



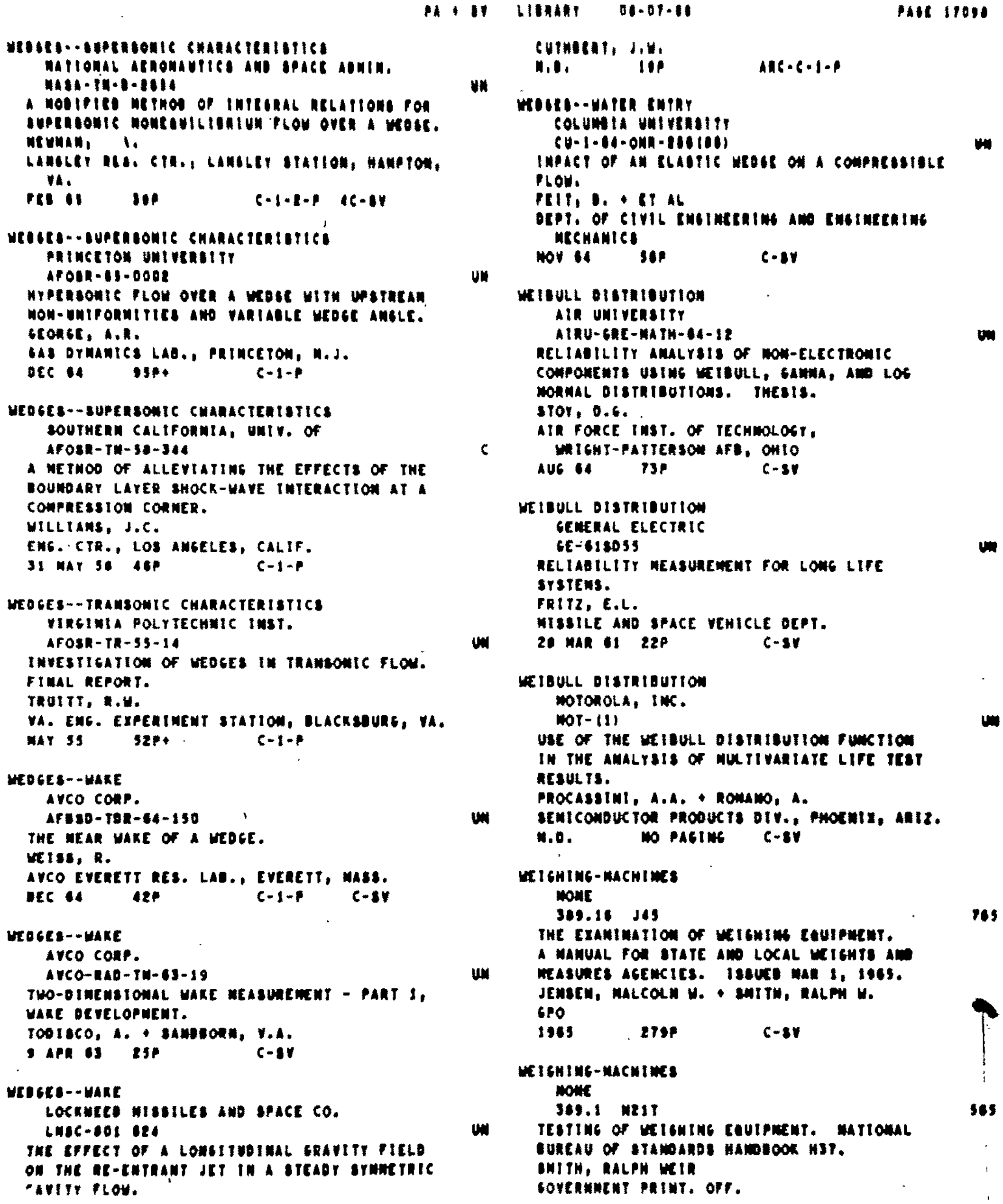
```
TDR 32-403                                   PA + SV   LIBRARY    08-07-66                    PAGE  4147

NEAL, J.T.
    AIR FORCE CAMBRIDGE RESEARCH LABS.
    AFCRL-65-375                                     UN
  GIANT DESICCATION POLYGONS OF GREAT BASIN
  PLAYAS.  ENVIRONMENTAL RESEARCH PAPERS,
  NO. 123.
  NEAL, J.T.
  BEDFORD, MASS.
  AUG 65      30P            C-1-P

NEAL, J.T., ED.
    AIR FORCE CAMBRIDGE RESEARCH LABS.
    AFCRL-65-266                                     UN
  GEOLOGY, MINERALOGY, AND HYDROLOGY OF U.S.
  PLAYAS.  ENVIRONMENTAL RESEARCH PAPERS,
  NO. 96.
  NEAL, J.T., ED.
  BEDFORD, MASS.
  APR 65      176P           C-1-P      2C-SV

NEAL, L.
    NATIONAL AERONAUTICS AND SPACE ADMIN.
    NASA-TM-X-816                                    C
  AN EXPLORATORY INVESTIGATION AT A MACH
  NUMBER OF 6.9 INTO THE USE OF AERODYNAMIC
  CONTROLS FOR MODULATING THE LIFT-DRAG RATIO
  OF AN APOLLO TYPE CONFIGURATION. (U)
  NEAL, L.
  LANGLEY RES. CTR., LANGLEY STATION,
    HAMPTON, VA.
  MAY 63      18P            C-1-2-P    4C-SV

NEAL, L.G.
    ARGONNE NATIONAL LAB.
    ANL-6625                                         UN
  LOCAL PARAMETERS IN COCURRENT MERCURY-
  NITROGEN FLOW.
  NEAK, L.G.
  ARGONNE, ILL.
  JAN 63      73P            C-SV

NEALE, D.M.
    NONE
    621.384  N25                             466
  COLD CATHODE TUBE CIRCUIT DESIGN.
  NEALE, D.M.
  PRINCETON, N.J., D. VAN NOSTRAND INC.
  1965        259P           SV

NEALE, L.C.
    WORCESTER POLYTECHNIC INST.
    WPL-AHL-49                                       UN
  REPORT ON EXPERIMENTAL INVESTIGATION OF
  CAVITATION ON BEHIND AN ACCELERATED DISC.
  NEALE, L.C. + STEVES, H.K.
  ALDEN HYDRAULIC LAB., WORCHESTER, MASS.
  FEB 60      13P            C-1-P

NEAVES, A.
    FRANKLIN INST.

    ASD-TDR-62-375                           UN
  RESEARCH ON SPONTANEOUS MAGNETIZATION IN
  SOLID BODIES.
  NEAVES, A.
  PHILADELPHIA, PA.
  APR 62      31P            PA

NEBIKER, F.R.
    GOODYEAR AEROSPACE CORP.
    FDL-TR-65-27                                     UN
  AERODYNAMIC DEPLOYABLE DECELERATOR
  PERFORMANCE-EVALUATION PROGRAM.
  NEBIKER, F.R.
  AKRON, OHIO
  AUG 65      305P           C-1-P

NEBLETTE, C.B.
    NONE
    771.35  N271                             366
  PHOTOGRAPHIC LENSES.
  NEBLETTE, C.B. + MURRAY, A.E.
  HASTINGS-ON-HUDSON, N.Y., MORGAN AND MORGAN
  1965        152P           SV

NECHAYEV, Y.N.
    FOREIGN TECHNOLOGY DIV.
    FTD-MT-64-301                                    UN
  AIR INTAKE DEVICES OF SUPERSONIC AIRCRAFT.
  NECHAYEV, Y.N.
  24 AUG 64  130P            C-SV

NECHELES, R.M.
    LOCKHEED-CALIFORNIA CO.
    LAC-LR-18759                                     UN(P)
  FULL LENGTH STRESS STRAIN CURVES AND UNIFORM
  ELONGATION MEASUREMENTS ON SELECTED TITANIUM
  ALLOYS - SST.
  NECHELES, R.M. + ET AL
  BURBANK, CALIF.
  5 APR 65   22P             C-SV

NECHELES, R.M.
    LOCKHEED-CALIFORNIA CO.
    LAC-LTM-50481                                    UN
  METALLURGICAL ANALYSIS OF FAILED BEARING
  TRUNNION NUT-MODEL 188.
  NECHELES, R.M.
  ENGINEERING RES. LAB., BURBANK, CALIF.
  19 AUG 63  VAR. PAGING  C-1-P

NECHELES, R.M.
    LOCKHEED-CALIFORNIA CO.
    LAC-LTM-50676                                    UN
  METALLURGICAL ANALYSIS OF SERVICE FAILED
  STABILIZER HOT LEADING EDGE - MODEL P-3A.
  NECHELES, R.M.
  ENGINEERING RES. LAB.
  2 MAR 64   VAR. PAGING  C-1-P

NECHELES, R.M.
```

FIG. 4. Authors



```
                              PA + SV   LIBRARY   08-07-66                PAGE   852

AF-33 (657)-11144
    STANFORD UNIVERSITY
    AFSC-AL-TDR-64-198                                C
  AN INSTANTANEOUS MICROWAVE POLARIMETER
  RECEIVER (U).  TECHNICAL REPORT NO. 1821-2.
  CRANE, R.
  STANFORD ELECTRONICS LAB., STANFORD, CALIF.
  MAY 64      66P             MICRO

AF-33 (657)-11144
    STANFORD UNIVERSITY
    AFSC-AL-TDR-64-227                                UN
  TUNING OF CW LASERS OVER ANGSTROM
  BANDWIDTHS - SOME POSSIBLE APPROACHES.
  MORRIS, R.J.
  AUG 64      40P             C-SV

AF-33 (657)-11154
    SYRACUSE UNIVERSITY
    AFSC-ML-TDR-64-144                                UN
  A STUDY OF THE EFFECT OF SUPERIMPOSED
  STRESS CONCENTRATIONS.
  WEISS, V. + ET AL
  SYRACUSE, N.Y.
  APR 64      34P             C-1-P

AF-33 (657)-11183
    AEROJET-GENERAL NUCLEONICS
    APL-TDR-64-124-VOL. I                             UN
  RESEARCH IN HIGH TEMPERATURE PLASMAS FOR
  SPACE-APPLICATIONS.  VOL. I - PHYSICS.
  NONE
  SAN RAMON, CALIF.
  OCT 64      180P            C-1-P

AF-33 (657)-11184
    MINNESOTA, UNIVERSITY OF
    FDL-TDR-64-156                                    UN
  A SECOND ORDER SOLUTION FOR THE VELOCITY
  DISTRIBUTION IN A TURBULENT WAKE.
  HEINRICH, H.G. + RUST, L.W.
  MINNEAPOLIS, MINN.
  APR 65      42P             C-1-P

AF-33 (657)-11200
    AEROJET-GENERAL CORP.
    AFSC-ML-TDR-64-260                                UN
  NON-EVACUATED CRYOGENIC THERMAL INSULATION
  STUDIES.
  JOHNSON, C.L. + HOLLWEGER, D.J.
  AZUSA, CALIF.
  SEPT 64     81P             C-1-P        C-SV

AF-33 (657)-11217
    AERONUTRONIC
    ASI-U-2577                                        S
  APPLICATION OF MATERIALS TO ADVANCED
  ROCKET NOZZLE AND HOT GAS CONTROL SYSTEMS.
  (U)  THIRD QUARTERLY PROGRESS REPORT.
  BLAES, N.M. + ET AL

  11 APR 64  VAR. PAGING  C-1-P(27)

AF-33 (657)-11223
    GENERAL ELECTRIC
    GE-11223-QPR-3                                    UN
  THE STRUCTURAL STABILITY OF WELDS IN
  COLUMBIUM ALLOYS.  PERIOD - NOV 1, 1963 -
  FEB 1, 1964.
  YOUNT, R.E. + KELLER, D.L.
  MATERIALS DEV. LAB. OPERATION, CINCINNATI,
    OHIO
  10 FEB 64  25P+           C-SV

AF-33 (657)-11233
    FRANKLIN INST.
    FRAN-I-B2122-1                                    UN
  DISTILLATION OF BERYLLIUM BY SUBLIMATION
  AND EVAPORATION.  INTERIM REPORT.  PERIOD -
  AUGUST 15 TO NOVEMBER 30, 1963.
  LONDON, G. + HERMAN. H.
  LABORATORIES FOR RES. AND DEV., PHILADELPHIA
    PA.
  N.D.        11P+          C-SV

AF-33 (657)-11253
    UNION CARBIDE CORP.
    AFSC-ML-TDR-64-173-PT. 3                          UN
  HIGH TEMPERATURE PROTECTIVE COATINGS FOR
  GRAPHITE.  PART -III.
  CRISCIONE, J.M. + ET AL
  PARMA, OHIO
  OCT 65      199P          C-1-P

AF-33 (657)-11316
    HONEYWELL, INC.
    FDL-TDR-64-69                                     UN
  TRAINABLE FLIGHT CONTROL SYSTEM
  INVESTIGATION.
  SMITH, F.B. + ET AL
  ST. PAUL, MINN.
  AUG 64      175P          C-1-P

AF-33 (657)-11326
    ELECTRO-OPTICAL SYSTEMS, INC.
    EOSI-3390-Q-1                                     UN
  OPTICALLY PUMPED IMAGE LIGHT AMPLIFICATION.
  QUARTERLY REPORT NO. 1.  PERIOD - 10 MAY -
  10 AUG 1963.
  BERNSTEIN, H. + ET AL
  PASADENA, CALIF.
  AUG 63      53P           C-SV

AF-33 (657)-11331
    SANTA RITA TECHNOLOGY, INC.
    MRL-TDR-63-60                                     UN
  AN ELECTRONIC ANALOG OF THE EAR.
  GLAESSER, E. + ET AL
  BIOACOUSTICS LAB. DIV., MENLO PARK, CALIF.
  JUNE 63     66P           C-1-P
```

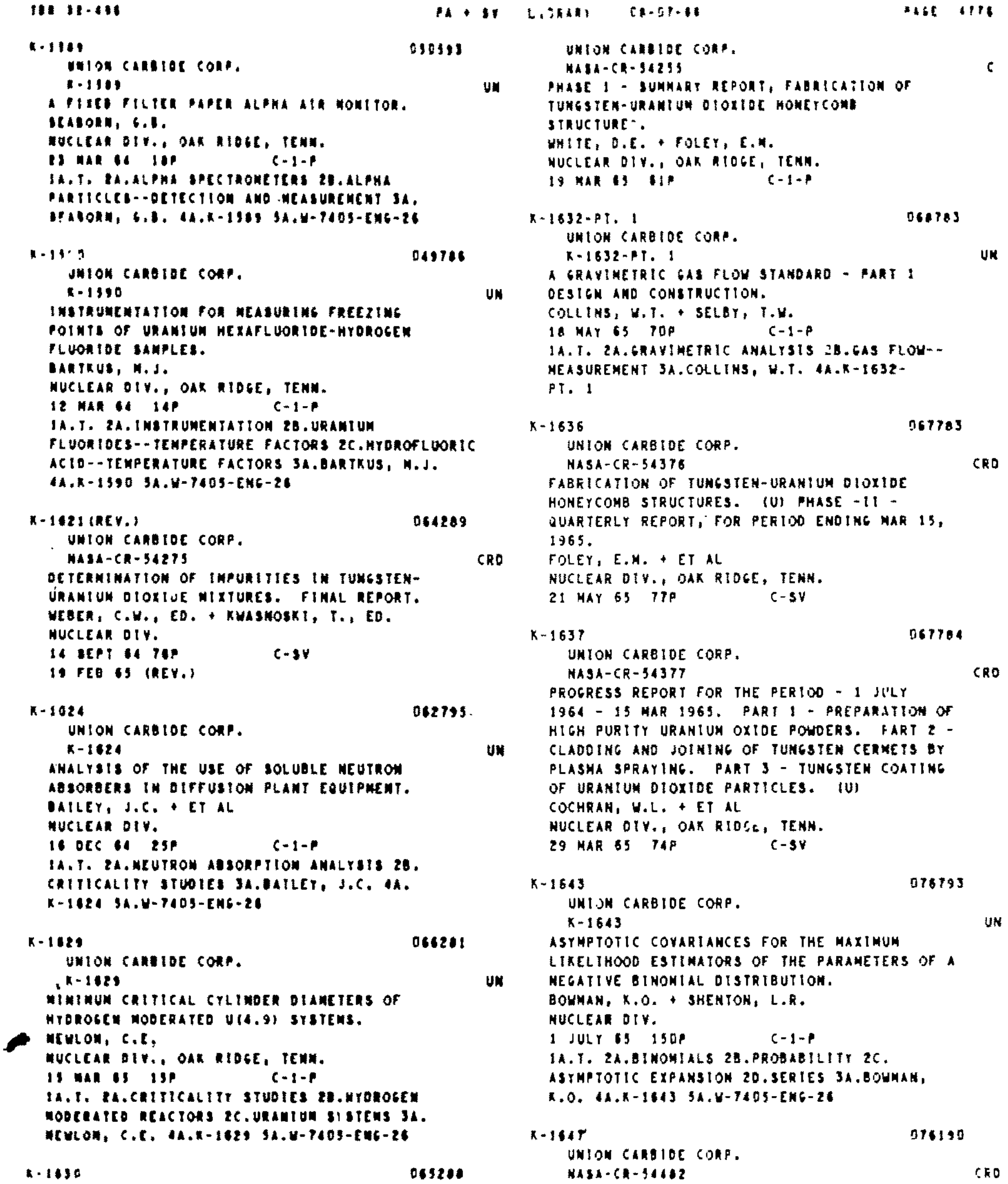
Fig. 5. Contract numbers



```
                                PA + SV   LIBRARY    06-07-66                        PAGE 17098

WEDGES--SUPERSONIC CHARACTERISTICS
    NATIONAL AERONAUTICS AND SPACE ADMIN.
    NASA-TN-D-2514                                    UN
  A MODIFIED METHOD OF INTEGRAL RELATIONS FOR
  SUPERSONIC NONEQUILIBRIUM FLOW OVER A WEDGE.
  NEWMAN,    \.
  LANGLEY RES. CTR.; LANGLEY STATION, HAMPTON,
    VA.
  FEB 65      30P           C-1-2-P  4C-SV

WEDGES--SUPERSONIC CHARACTERISTICS
    PRINCETON UNIVERSITY
    AFOSR-65-0002                                     UN
  HYPERSONIC FLOW OVER A WEDGE WITH UPSTREAM
  NON-UNIFORMITIES AND VARIABLE WEDGE ANGLE.
  GEORGE, A.R.
  GAS DYNAMICS LAB., PRINCETON, N.J.
  DEC 64      95P+          C-1-P

WEDGES--SUPERSONIC CHARACTERISTICS
    SOUTHERN CALIFORNIA, UNIV. OF
    AFOSR-TN-58-344                                    C
  A METHOD OF ALLEVIATING THE EFFECTS OF THE
  BOUNDARY LAYER SHOCK-WAVE INTERACTION AT A
  COMPRESSION CORNER.
  WILLIAMS, J.C.
  ENG. CTR., LOS ANGELES, CALIF.
  31 MAY 58  46P            C-1-P

WEDGES--TRANSONIC CHARACTERISTICS
    VIRGINIA POLYTECHNIC INST.
    AFOSR-TR-55-14                                    UN
  INVESTIGATION OF WEDGES IN TRANSONIC FLOW.
  FINAL REPORT.
  TRUITT, R.W.
  VA. ENG. EXPERIMENT STATION, BLACKSBURG, VA.
  MAY 55      52P+          C-1-P

WEDGES--WAKE
    AVCO CORP.
    AFSSD-TDR-64-150                                  UN
  THE NEAR WAKE OF A WEDGE.
  WEISS, R.
  AVCO EVERETT RES. LAB., EVERETT, MASS.
  DEC 64      42P           C-1-P     C-SV

WEDGES--WAKE
    AVCO CORP.
    AVCO-RAD-TM-63-19                                 UN
  TWO-DIMENSIONAL WAKE MEASUREMENT - PART 1,
  WAKE DEVELOPMENT.
  TODISCO, A. + SANDBORN, V.A.
  9 APR 63    25P           C-SV

WEDGES--WAKE
    LOCKHEED MISSILES AND SPACE CO.
    LMSC-801 624                                      UN
  THE EFFECT OF A LONGITUDINAL GRAVITY FIELD
  ON THE RE-ENTRANT JET IN A STEADY SYMMETRIC
  "AVITY FLOW.
```

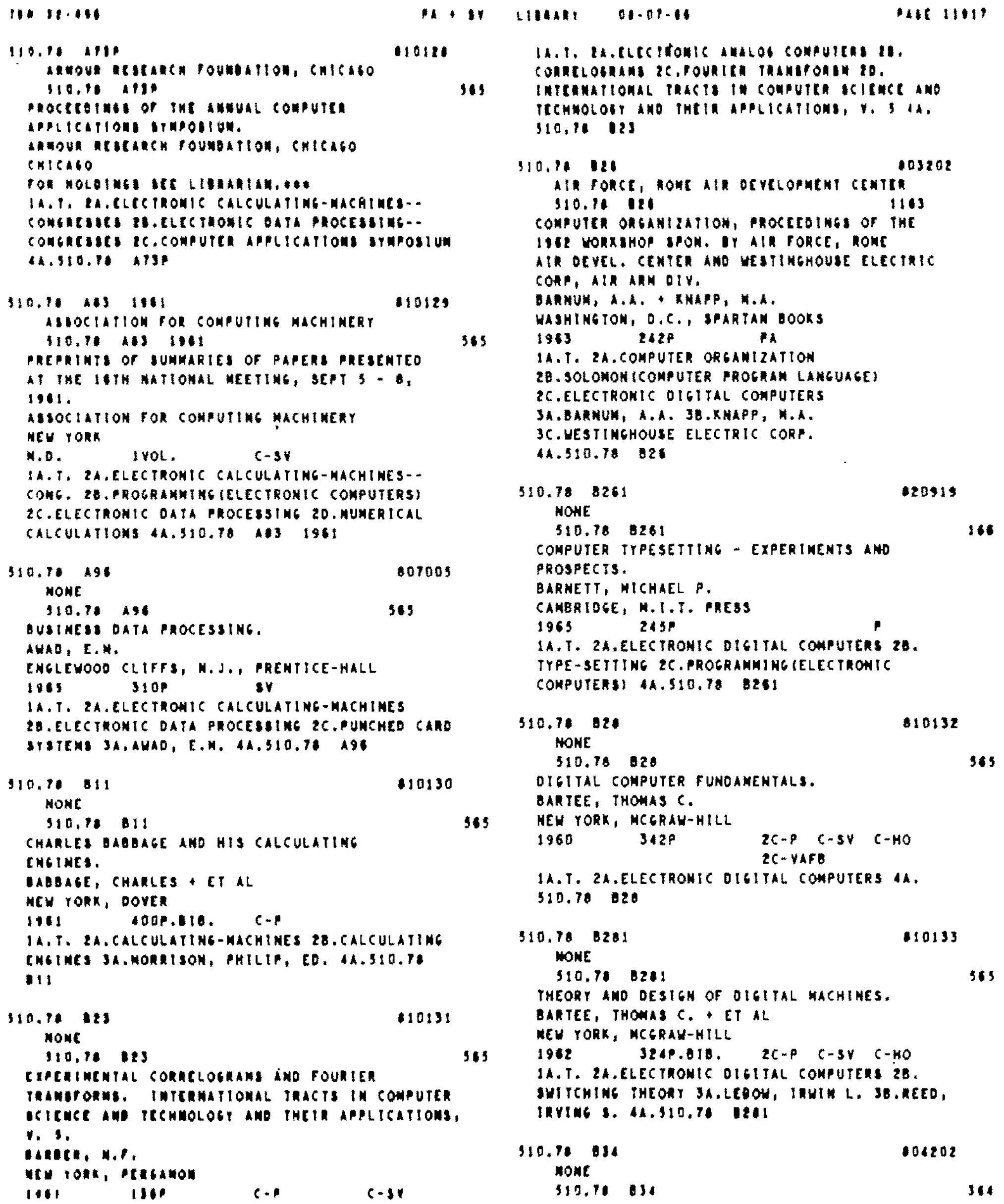
```
  CUTHBERT, J.W.
  N.D.        10P           ARC-C-1-P

WEDGES--WATER ENTRY
    COLUMBIA UNIVERSITY
    CU-1-64-ONR-266(08)                                      UN
  IMPACT OF AN ELASTIC WEDGE ON A COMPRESSIBLE
  FLOW.
  FEIT, D. + ET AL
  DEPT. OF CIVIL ENGINEERING AND ENGINEERING
    MECHANICS
  NOV 64      58P           C-SV

WEIBULL DISTRIBUTION
    AIR UNIVERSITY
    AIRU-GRE-MATH-64-12                                      UN
  RELIABILITY ANALYSIS OF NON-ELECTRONIC
  COMPONENTS USING WEIBULL, GAMMA, AND LOG
  NORMAL DISTRIBUTIONS.  THESIS.
  STOY, D.G.
  AIR FORCE INST. OF TECHNOLOGY,
    WRIGHT-PATTERSON AFB, OHIO
  AUG 64      73P           C-SV

WEIBULL DISTRIBUTION
    GENERAL ELECTRIC
    GE-61SD55                                                UN
  RELIABILITY MEASUREMENT FOR LONG LIFE
  SYSTEMS.
  FRITZ, E.L.
  MISSILE AND SPACE VEHICLE DEPT.
  20 MAR 61  22P            C-SV

WEIBULL DISTRIBUTION
    MOTOROLA, INC.
    MOT-(1)                                                  UN
  USE OF THE WEIBULL DISTRIBUTION FUNCTION
  IN THE ANALYSIS OF MULTIVARIATE LIFE TEST
  RESULTS.
  PROCASSINI, A.A. + ROMANO, A.
  SEMICONDUCTOR PRODUCTS DIV., PHOENIX, ARIZ.
  N.D.        NO PAGING     C-SV

WEIGHING-MACHINES
    NONE
    389.16  J45                                              765
  THE EXAMINATION OF WEIGHING EQUIPMENT.
  A MANUAL FOR STATE AND LOCAL WEIGHTS AND
  MEASURES AGENCIES.  ISSUED MAR 1, 1965.
  JENSEN, MALCOLM W. + SMITH, RALPH W.
  GPO
  1965        279P          C-SV

WEIGHING-MACHINES
    NONE
    389.1  N21T                                              565
  TESTING OF WEIGHING EQUIPMENT.  NATIONAL
  BUREAU OF STANDARDS HANDBOOK H37.
  SMITH, RALPH WEIR
  GOVERNMENT PRINT. OFF.
```

FIG. 6. Subjects



```
TID 38-498                                    PA + SV   LIBRARY    CR-67-08                 PAGE  4776

K-1589                             050593           UNION CARBIDE CORP.
    UNION CARBIDE CORP.                             NASA-CR-54255                          C
    K-1589                             UN        PHASE 1 - SUMMARY REPORT, FABRICATION OF
  A FIXED FILTER PAPER ALPHA AIR MONITOR.        TUNGSTEN-URANIUM DIOXIDE HONEYCOMB
  SEABORN, G.B.                                  STRUCTURE".
  NUCLEAR DIV., OAK RIDGE, TENN.                 WHITE, D.E. + FOLEY, E.M.
  23 MAR 64  18P           C-1-P                 NUCLEAR DIV., OAK RIDGE, TENN.
  1A.T. 2A.ALPHA SPECTROMETERS 2B.ALPHA          19 MAR 65  81P          C-1-P
  PARTICLES--DETECTION AND MEASUREMENT 3A.
  SEABORN, G.B. 4A.K-1589 5A.W-7405-ENG-26     K-1632-PT. 1                        068783
                                                    UNION CARBIDE CORP.
K-1590                             049786           K-1632-PT. 1                           UN
    UNION CARBIDE CORP.                          A GRAVIMETRIC GAS FLOW STANDARD - PART 1
    K-1590                             UN        DESIGN AND CONSTRUCTION.
  INSTRUMENTATION FOR MEASURING FREEZING         COLLINS, W.T. + SELBY, T.W.
  POINTS OF URANIUM HEXAFLUORIDE-HYDROGEN        18 MAY 65  70P          C-1-P
  FLUORIDE SAMPLES.                              1A.T. 2A.GRAVIMETRIC ANALYSIS 2B.GAS FLOW--
  BARTKUS, M.J.                                  MEASUREMENT 3A.COLLINS, W.T. 4A.K-1632-
  NUCLEAR DIV., OAK RIDGE, TENN.                 PT. 1
  12 MAR 64  14P           C-1-P
  1A.T. 2A.INSTRUMENTATION 2B.URANIUM          K-1636                              067783
  FLUORIDES--TEMPERATURE FACTORS 2C.HYDROFLUORIC    UNION CARBIDE CORP.
  ACID--TEMPERATURE FACTORS 3A.BARTKUS, M.J.        NASA-CR-54376                          CRD
  4A.K-1590 5A.W-7405-ENG-26                     FABRICATION OF TUNGSTEN-URANIUM DIOXIDE
                                                 HONEYCOMB STRUCTURES.  (U) PHASE -II -
K-1621(REV.)                       064289        QUARTERLY REPORT, FOR PERIOD ENDING MAR 15,
    UNION CARBIDE CORP.                          1965.
    NASA-CR-54275                      CRD       FOLEY, E.M. + ET AL
  DETERMINATION OF IMPURITIES IN TUNGSTEN-       NUCLEAR DIV., OAK RIDGE, TENN.
  URANIUM DIOXIJE MIXTURES.  FINAL REPORT.       21 MAY 65  77P          C-SV
  WEBER, C.W., ED. + KWASNOSKI, T., ED.
  NUCLEAR DIV.                                 K-1637                              067784
  14 SEPT 64 78P           C-SV                     UNION CARBIDE CORP.
  19 FEB 65 (REV.)                                  NASA-CR-54377                          CRD
                                                 PROGRESS REPORT FOR THE PERIOD - 1 JULY
K-1624                             062795.       1964 - 15 MAR 1965.  PART 1 - PREPARATION OF
    UNION CARBIDE CORP.                          HIGH PURITY URANIUM OXIDE POWDERS.  PART 2 -
    K-1624                             UN        CLADDING AND JOINING OF TUNGSTEN CERMETS BY
  ANALYSIS OF THE USE OF SOLUBLE NEUTRON         PLASMA SPRAYING.  PART 3 - TUNGSTEN COATING
  ABSORBERS IN DIFFUSION PLANT EQUIPMENT.        OF URANIUM DIOXIDE PARTICLES.  (U)
  BAILEY, J.C. + ET AL                           COCHRAN, W.L. + ET AL
  NUCLEAR DIV.                                   NUCLEAR DIV., OAK RIDGE, TENN.
  16 DEC 64  25P           C-1-P                 29 MAR 65  74P          C-SV
  1A.T. 2A.NEUTRON ABSORPTION ANALYSIS 2B.
  CRITICALITY STUDIES 3A.BAILEY, J.C. 4A.      K-1643                              076793
  K-1624 5A.W-7405-ENG-26                           UNION CARBIDE CORP.
                                                    K-1643                                 UN
K-1629                             066281        ASYMPTOTIC COVARIANCES FOR THE MAXIMUM
    UNION CARBIDE CORP.                          LIKELIHOOD ESTIMATORS OF THE PARAMETERS OF A
    K-1629                             UN        NEGATIVE BINOMIAL DISTRIBUTION.
  MINIMUM CRITICAL CYLINDER DIAMETERS OF         BOWMAN, K.O. + SHENTON, L.R.
  HYDROGEN MODERATED U(4.9) SYSTEMS.             NUCLEAR DIV.
  NEWLON, C.E.                                   1 JULY 65  150P         C-1-P
  NUCLEAR DIV., OAK RIDGE, TENN.                 1A.T. 2A.BINOMIALS 2B.PROBABILITY 2C.
  15 MAR 65  15P           C-1-P                 ASYMPTOTIC EXPANSION 2D.SERIES 3A.BOWMAN,
  1A.T. 2A.CRITICALITY STUDIES 2B.HYDROGEN       K.O. 4A.K-1643 5A.W-7405-ENG-26
  MODERATED REACTORS 2C.URANIUM SISTEMS 3A.
  NEWLON, C.E. 4A.K-1629 5A.W-7405-ENG-26      K-1647                              076190
                                                    UNION CARBIDE CORP.
K-1630                             065280           NASA-CR-54482                          CRD
```

FIG. 7. Report numbers



```
TRW 22-496                                 PA + SV

510.78  A73P                                810128
   ARMOUR RESEARCH FOUNDATION, CHICAGO
   510.78  A73P                                      565
  PROCEEDINGS OF THE ANNUAL COMPUTER
  APPLICATIONS SYMPOSIUM.
  ARMOUR RESEARCH FOUNDATION, CHICAGO
  CHICAGO
  FOR HOLDINGS SEE LIBRARIAN.***
  1A.T. 2A.ELECTRONIC CALCULATING-MACHINES--
  CONGRESSES 2B.ELECTRONIC DATA PROCESSING--
  CONGRESSES 2C.COMPUTER APPLICATIONS SYMPOSIUM
  4A.510.78  A73P

510.78  A83  1961                           810129
   ASSOCIATION FOR COMPUTING MACHINERY
   510.78  A83  1961                                 565
  PREPRINTS OF SUMMARIES OF PAPERS PRESENTED
  AT THE 16TH NATIONAL MEETING, SEPT 5 - 8,
  1961.
  ASSOCIATION FOR COMPUTING MACHINERY
  NEW YORK
  N.D.          1VOL.          C-SV
  1A.T. 2A.ELECTRONIC CALCULATING-MACHINES--
  CONG. 2B.PROGRAMMING(ELECTRONIC COMPUTERS)
  2C.ELECTRONIC DATA PROCESSING 2D.NUMERICAL
  CALCULATIONS 4A.510.78  A83  1961

510.78  A96                                 807005
   NONE
   510.78  A96                            565
  BUSINESS DATA PROCESSING.
  AWAD, E.M.
  ENGLEWOOD CLIFFS, N.J., PRENTICE-HALL
  1965          510P           SV
  1A.T. 2A.ELECTRONIC CALCULATING-MACHINES
  2B.ELECTRONIC DATA PROCESSING 2C.PUNCHED CARD
  SYSTEMS 3A.AWAD, E.M. 4A.510.78  A96

510.78  B11                                 810130
   NONE
   510.78  B11                                       565
  CHARLES BABBAGE AND HIS CALCULATING
  ENGINES.
  BABBAGE, CHARLES + ET AL
  NEW YORK, DOVER
  1961          400P.BIB.      C-P
  1A.T. 2A.CALCULATING-MACHINES 2B.CALCULATING
  ENGINES 3A.MORRISON, PHILIP, ED. 4A.510.78
  B11

510.78  B23                                 810131
   NONE
   510.78  B23                                       565
  EXPERIMENTAL CORRELOGRAMS AND FOURIER
  TRANSFORMS.  INTERNATIONAL TRACTS IN COMPUTER
  SCIENCE AND TECHNOLOGY AND THEIR APPLICATIONS,
  V. 5.
  BARBER, N.F.
  NEW YORK, PERGAMON
  1961          136P           C-P            C-SV
```

```
LIBRARY     08-07-66                         PAGE 11017

  1A.T. 2A.ELECTRONIC ANALOG COMPUTERS 2B.
  CORRELOGRAMS 2C.FOURIER TRANSFORMS 2D.
  INTERNATIONAL TRACTS IN COMPUTER SCIENCE AND
  TECHNOLOGY AND THEIR APPLICATIONS, V. 5 4A.
  510.78  B23

510.78  B26                                 803202
   AIR FORCE, ROME AIR DEVELOPMENT CENTER
   510.78  B26                              1163
  COMPUTER ORGANIZATION, PROCEEDINGS OF THE
  1962 WORKSHOP SPON. BY AIR FORCE, ROME
  AIR DEVEL. CENTER AND WESTINGHOUSE ELECTRIC
  CORP, AIR ARM DIV.
  BARNUM, A.A. + KNAPP, M.A.
  WASHINGTON, D.C., SPARTAN BOOKS
  1963          242P           PA
  1A.T. 2A.COMPUTER ORGANIZATION
  2B.SOLOMON(COMPUTER PROGRAM LANGUAGE)
  2C.ELECTRONIC DIGITAL COMPUTERS
  3A.BARNUM, A.A. 3B.KNAPP, M.A.
  3C.WESTINGHOUSE ELECTRIC CORP.
  4A.510.78  B26

510.78  B261                                820919
   NONE
   510.78  B261                                      166
  COMPUTER TYPESETTING - EXPERIMENTS AND
  PROSPECTS.
  BARNETT, MICHAEL P.
  CAMBRIDGE, M.I.T. PRESS
  1965          245P                        P
  1A.T. 2A.ELECTRONIC DIGITAL COMPUTERS 2B.
  TYPE-SETTING 2C.PROGRAMMING(ELECTRONIC
  COMPUTERS) 4A.510.78  B261

510.78  B28                                 810132
   NONE
   510.78  B28                                       565
  DIGITAL COMPUTER FUNDAMENTALS.
  BARTEE, THOMAS C.
  NEW YORK, MCGRAW-HILL
  1960          342P           2C-P  C-SV  C-MO
                               2C-VAFB
  1A.T. 2A.ELECTRONIC DIGITAL COMPUTERS 4A.
  510.78  B28

510.78  B281                                810133
   NONE
   510.78  B281                                      565
  THEORY AND DESIGN OF DIGITAL MACHINES.
  BARTEE, THOMAS C. + ET AL
  NEW YORK, MCGRAW-HILL
  1962          324P.BIB.      2C-P  C-SV  C-MO
  1A.T. 2A.ELECTRONIC DIGITAL COMPUTERS 2B.
  SWITCHING THEORY 3A.LEBOW, IRWIN L. 3B.REED,
  IRVING S. 4A.510.78  B281

510.78  B34                                 804202
   NONE
   510.78  B34                                       364
```

FIG. 8. Book call numbers

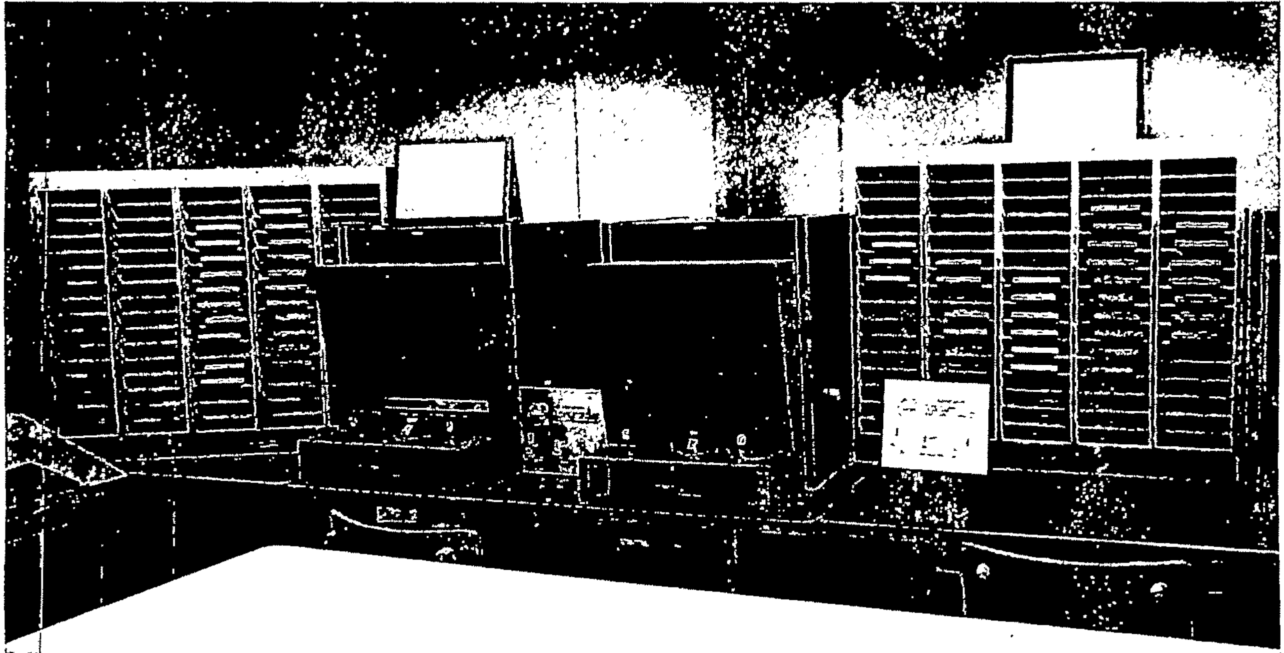


FIG. 9. Microfilm catalog installation

The computer-produced microfilm catalog provides bonuses that are immediately attractive to librarians. Perhaps the chief bonus is a separate catalog with reader for the library's technical services staff. This point is the more important at LMSC because its TIC has separate staffs for reports and for book acquisitions and cataloging. Having the complete catalog in their own work stations obviously permits the technical services librarians to function more efficiently.

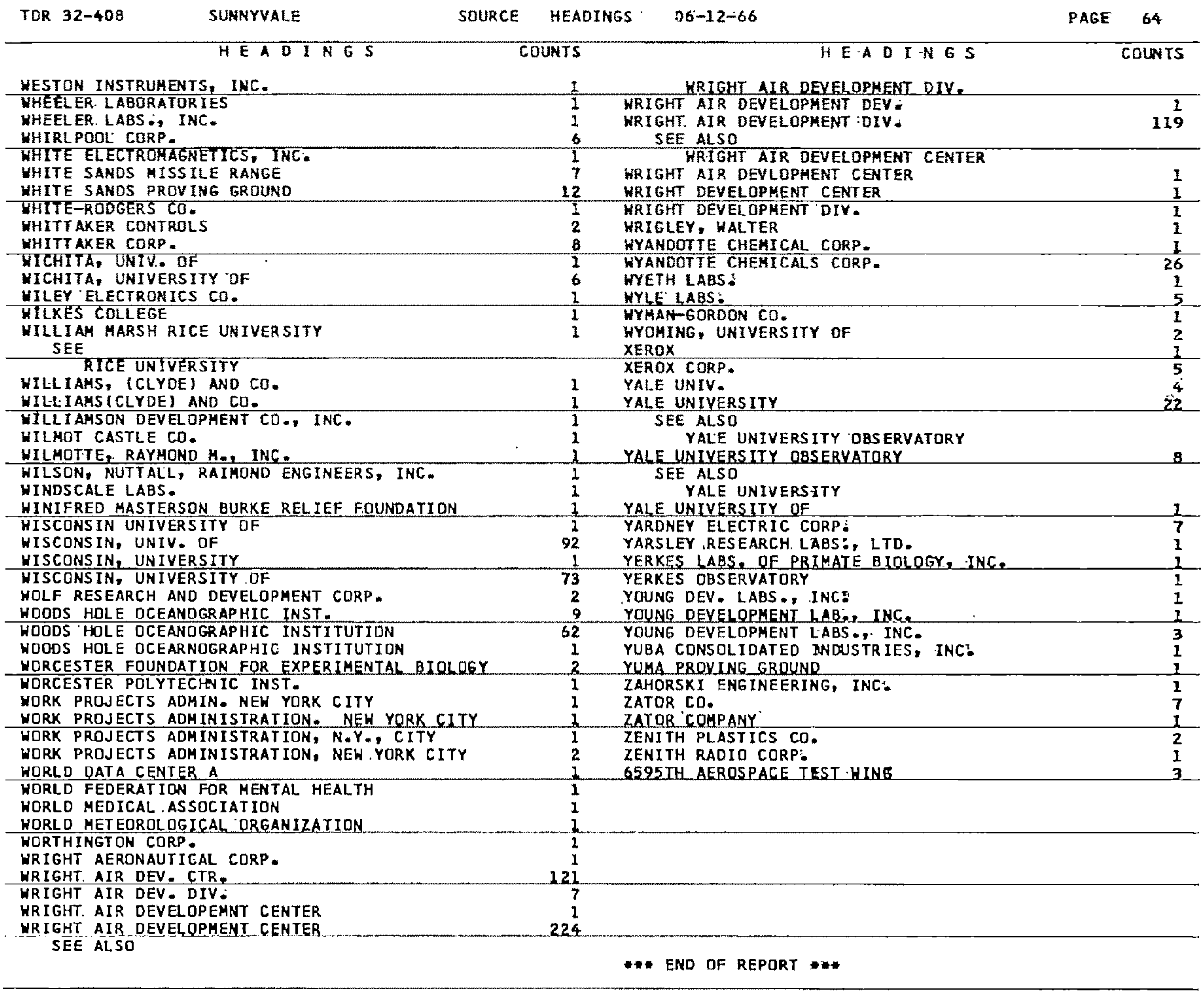
Similar bonuses that have equal beneficial effects on efficiency of library operations are the separate catalog/reader installations for the literature search corps and the reports and books circulation desks at one of the TIC's two libraries; namely, the one that serves approximately 11,051 scientists, engineers, and administrative support personnel. The needs of literature search are self-evident, and the volume of loan requests handled by these two service desks as well as their distance from the public catalog installation made it operationally and economically feasible to install these catalogs.

Additional bonuses generated by this computer-produced microfilm library catalog system are: (1) authority list of sources; (2) authority list of subject headings; and (3) list of open-entry items. While these lists are products of the system's specifications, they can be identified as bonuses in relation to their nonexistence under the displaced system. There is no need to stress the operational importance of the first two lists (Figs. 10 and 11), especially since they are automatically updated to reflect professional decisions of deletion and addition as well as of augmentation by integrative cross-referencing. Operational utility is enhanced by printing sufficient copies to supply a set to each cataloger and a set for the literature search corps. The third list (Fig. 12) identifies the TIC's holdings of cataloged "serial" titles. Included in this concept are reports generated on a specific contract, project, task, or other effort which are uniquely identifiable by the same report number in extension, e.g., LMSC 1481-2, LMSC 1481-3, etc. Such report titles are listed but once in the official microfilm catalog and then with the notation: "See librarian for holdings." The librarian consults her open-entry list and satisfies the requestor as to holdings.

The most far-reaching spin-off of the computer-produced microfilm library catalog system, however, is its power to deliver a printed book catalog at exceptionally low costs. The savings reside chiefly in the absence of photographic expenses and of press set-up costs. The library administrator, whose clientele would object to using a microfilm catalog, could use the computer-produced microfilm system as the printing base for his book catalog since it cuts printing costs by two-thirds (Table 1).

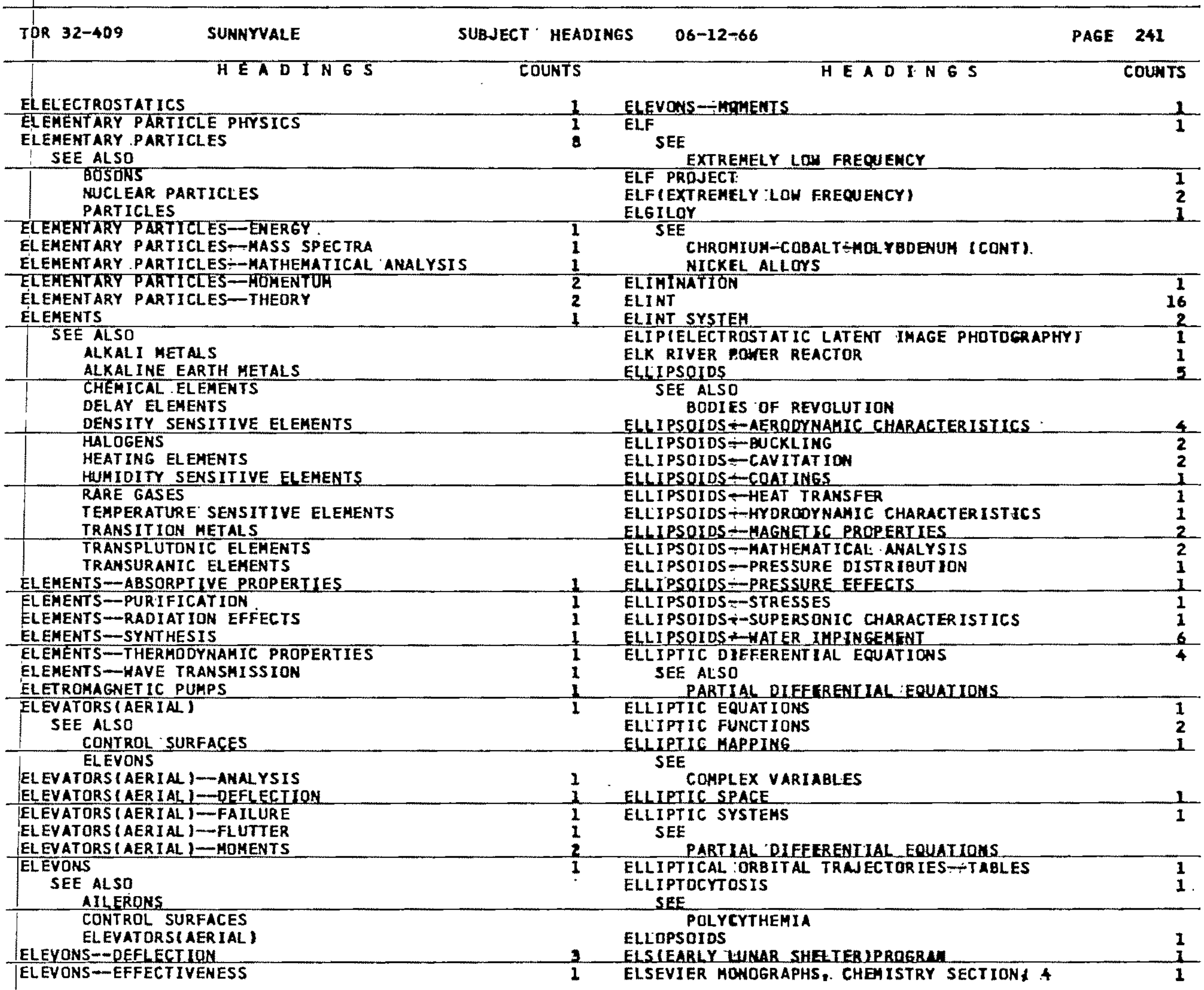
TABLE 1. Comparative printing costs of library book catalog processed, *A*, from microfilm master (using Copyflo process) and, *B*, from computer print-out (using ITEK process)—based on 1,800 pages printed head-to-head, simple binding, each volume 300 pages

Operation	A Copyflo		B Multilith Printing	
	20 Copies	1 Copy	20 Copies	1 Copy
Plates	$136.80	—	$ 642.60	$ 642.60
Press Set Up	—	—	343.80	343.80
Bond Paper	—	$73.80	—	—
Press Run	216.00	—	—	—
Impressions	—	—	180.00	9.00
Collation	36.00	3.60	36.00	3.60
Binding	20.80	4.16	20.80	2.08
Total	$409.60	$81.56	$1223.20	$1001.08



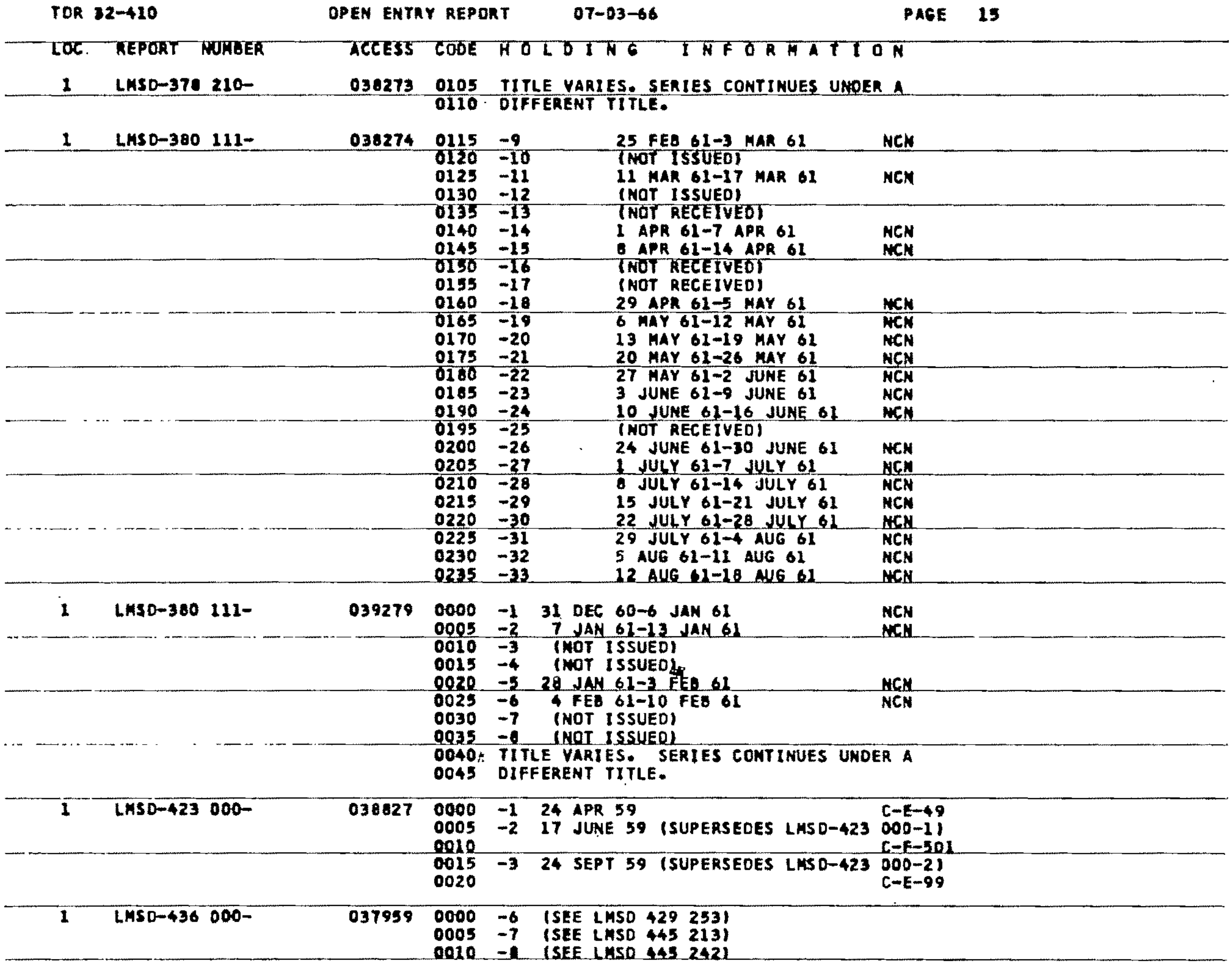
```
TDR 32-408         SUNNYVALE               SOURCE   HEADINGS   06-12-66                                  PAGE  64

               H E A D I N G S                      COUNTS                   H E A D I N G S                    COUNTS

WESTON INSTRUMENTS, INC.                                1              WRIGHT AIR DEVELOPMENT DIV.
WHEELER LABORATORIES                                    1      WRIGHT AIR DEVELOPMENT DEV.                           1
WHEELER LABS., INC.                                     1      WRIGHT AIR DEVELOPMENT DIV.                         119
WHIRLPOOL CORP.                                         6         SEE ALSO
WHITE ELECTROMAGNETICS, INC.                            1              WRIGHT AIR DEVELOPMENT CENTER
WHITE SANDS MISSILE RANGE                               7      WRIGHT AIR DEVLOPMENT CENTER                          1
WHITE SANDS PROVING GROUND                             12      WRIGHT DEVELOPMENT CENTER                             1
WHITE-RODGERS CO.                                       1      WRIGHT DEVELOPMENT DIV.                               1
WHITTAKER CONTROLS                                      2      WRIGLEY, WALTER                                       1
WHITTAKER CORP.                                         8      WYANDOTTE CHEMICAL CORP.                              1
WICHITA, UNIV. OF                                       1      WYANDOTTE CHEMICALS CORP.                            26
WICHITA, UNIVERSITY OF                                  6      WYETH LABS.                                           1
WILEY ELECTRONICS CO.                                   1      WYLE LABS.                                            5
WILKES COLLEGE                                          1      WYMAN-GORDON CO.                                      1
WILLIAM MARSH RICE UNIVERSITY                           1      WYOMING, UNIVERSITY OF                                2
   SEE                                                         XEROX                                                 1
      RICE UNIVERSITY                                          XEROX CORP.                                           5
WILLIAMS, (CLYDE) AND CO.                               1      YALE UNIV.                                            4
WILLIAMS(CLYDE) AND CO.                                 1      YALE UNIVERSITY                                      22
WILLIAMSON DEVELOPMENT CO., INC.                        1         SEE ALSO
WILMOT CASTLE CO.                                       1              YALE UNIVERSITY OBSERVATORY
WILMOTTE, RAYMOND M., INC.                              1      YALE UNIVERSITY OBSERVATORY                           8
WILSON, NUTTALL, RAIMOND ENGINEERS, INC.                1         SEE ALSO
WINDSCALE LABS.                                         1              YALE UNIVERSITY
WINIFRED MASTERSON BURKE RELIEF FOUNDATION              1      YALE UNIVERSITY OF                                    1
WISCONSIN UNIVERSITY OF                                 1      YARDNEY ELECTRIC CORP.                                7
WISCONSIN, UNIV. OF                                    92      YARSLEY RESEARCH LABS., LTD.                          1
WISCONSIN, UNIVERSITY                                   1      YERKES LABS. OF PRIMATE BIOLOGY, INC.                 1
WISCONSIN, UNIVERSITY OF                               73      YERKES OBSERVATORY                                    1
WOLF RESEARCH AND DEVELOPMENT CORP.                     2      YOUNG DEV. LABS., INC.                                1
WOODS HOLE OCEANOGRAPHIC INST.                          9      YOUNG DEVELOPMENT LAB., INC.                          1
WOODS HOLE OCEANOGRAPHIC INSTITUTION                   62      YOUNG DEVELOPMENT LABS., INC.                         3
WOODS HOLE OCEARNOGRAPHIC INSTITUTION                   1      YUBA CONSOLIDATED INDUSTRIES, INC.                    1
WORCESTER FOUNDATION FOR EXPERIMENTAL BIOLOGY           2      YUMA PROVING GROUND                                   1
WORCESTER POLYTECHNIC INST.                             1      ZAHORSKI ENGINEERING, INC.                            1
WORK PROJECTS ADMIN. NEW YORK CITY                      1      ZATOR CO.                                             7
WORK PROJECTS ADMINISTRATION.  NEW YORK CITY            1      ZATOR COMPANY                                         1
WORK PROJECTS ADMINISTRATION, N.Y., CITY                1      ZENITH PLASTICS CO.                                   2
WORK PROJECTS ADMINISTRATION, NEW YORK CITY             2      ZENITH RADIO CORP.                                    1
WORLD DATA CENTER A                                     1      6595TH AEROSPACE TEST WING                            3
WORLD FEDERATION FOR MENTAL HEALTH                      1
WORLD MEDICAL ASSOCIATION                               1
WORLD METEOROLOGICAL ORGANIZATION                       1
WORTHINGTON CORP.                                       1
WRIGHT AERONAUTICAL CORP.                               1
WRIGHT AIR DEV. CTR.                                  121
WRIGHT AIR DEV. DIV.                                    7
WRIGHT AIR DEVELOPEMNT CENTER                           1
WRIGHT AIR DEVELOPMENT CENTER                         224
   SEE ALSO
                                                               *** END OF REPORT ***
```

Fig. 10. Source authority list



```
TDR 32-409          SUNNYVALE              SUBJECT HEADINGS    06-12-66                      PAGE  241

                 H E A D I N G S           COUNTS                 H E A D I N G S            COUNTS

ELELECTROSTATICS                               1    ELEVONS--MOMENTS                               1
ELEMENTARY PARTICLE PHYSICS                    1    ELF                                            1
ELEMENTARY PARTICLES                           8      SEE
  SEE ALSO                                              EXTREMELY LOW FREQUENCY
      BOSONS                                        ELF PROJECT                                    1
      NUCLEAR PARTICLES                             ELF(EXTREMELY LOW FREQUENCY)                   2
      PARTICLES                                     ELGILOY                                        1
ELEMENTARY PARTICLES--ENERGY                   1      SEE
ELEMENTARY PARTICLES--MASS SPECTRA             1        CHROMIUM-COBALT-MOLYBDENUM (CONT).
ELEMENTARY PARTICLES--MATHEMATICAL ANALYSIS    1        NICKEL ALLOYS
ELEMENTARY PARTICLES--MOMENTUM                 2    ELIMINATION                                    1
ELEMENTARY PARTICLES--THEORY                   2    ELINT                                         16
ELEMENTS                                       1    ELINT SYSTEM                                   2
  SEE ALSO                                          ELIP(ELECTROSTATIC LATENT IMAGE PHOTOGRAPHY)   1
      ALKALI METALS                                 ELK RIVER POWER REACTOR                        1
      ALKALINE EARTH METALS                         ELLIPSOIDS                                     5
      CHEMICAL ELEMENTS                               SEE ALSO
      DELAY ELEMENTS                                    BODIES OF REVOLUTION
      DENSITY SENSITIVE ELEMENTS                    ELLIPSOIDS--AERODYNAMIC CHARACTERISTICS        4
      HALOGENS                                      ELLIPSOIDS--BUCKLING                           2
      HEATING ELEMENTS                              ELLIPSOIDS--CAVITATION                         2
      HUMIDITY SENSITIVE ELEMENTS                   ELLIPSOIDS--COATINGS                           1
      RARE GASES                                    ELLIPSOIDS--HEAT TRANSFER                      1
      TEMPERATURE SENSITIVE ELEMENTS                ELLIPSOIDS--HYDRODYNAMIC CHARACTERISTICS       1
      TRANSITION METALS                             ELLIPSOIDS--MAGNETIC PROPERTIES                2
      TRANSPLUTONIC ELEMENTS                        ELLIPSOIDS--MATHEMATICAL ANALYSIS              2
      TRANSURANIC ELEMENTS                          ELLIPSOIDS--PRESSURE DISTRIBUTION              1
ELEMENTS--ABSORPTIVE PROPERTIES                1    ELLIPSOIDS--PRESSURE EFFECTS                   1
ELEMENTS--PURIFICATION                         1    ELLIPSOIDS--STRESSES                           1
ELEMENTS--RADIATION EFFECTS                    1    ELLIPSOIDS--SUPERSONIC CHARACTERISTICS         1
ELEMENTS--SYNTHESIS                            1    ELLIPSOIDS--WATER IMPINGEMENT                  6
ELEMENTS--THERMODYNAMIC PROPERTIES             1    ELLIPTIC DIFFERENTIAL EQUATIONS                4
ELEMENTS--WAVE TRANSMISSION                    1      SEE ALSO
ELETROMAGNETIC PUMPS                           1        PARTIAL DIFFERENTIAL EQUATIONS
ELEVATORS(AERIAL)                              1    ELLIPTIC EQUATIONS                             1
  SEE ALSO                                          ELLIPTIC FUNCTIONS                             2
      CONTROL SURFACES                              ELLIPTIC MAPPING                               1
      ELEVONS                                         SEE
ELEVATORS(AERIAL)--ANALYSIS                    1        COMPLEX VARIABLES
ELEVATORS(AERIAL)--DEFLECTION                  1    ELLIPTIC SPACE                                 1
ELEVATORS(AERIAL)--FAILURE                     1    ELLIPTIC SYSTEMS                               1
ELEVATORS(AERIAL)--FLUTTER                     1      SEE
ELEVATORS(AERIAL)--MOMENTS                     2        PARTIAL DIFFERENTIAL EQUATIONS
ELEVONS                                        1    ELLIPTICAL ORBITAL TRAJECTORIES--TABLES        1
  SEE ALSO                                          ELLIPTOCYTOSIS                                 1
      AILERONS                                        SEE
      CONTROL SURFACES                                  POLYCYTHEMIA
      ELEVATORS(AERIAL)                             ELLOPSOIDS                                     1
ELEVONS--DEFLECTION                            3    ELS(EARLY LUNAR SHELTER)PROGRAM                1
ELEVONS--EFFECTIVENESS                         1    ELSEVIER MONOGRAPHS, CHEMISTRY SECTION, 4      1
```

FIG. 11. Subject authority list



```
TDR 32-410              OPEN ENTRY REPORT      07-03-66                  PAGE   15

LOC.  REPORT  NUMBER        ACCESS  CODE  H O L D I N G   I N F O R M A T I O N

 1    LMSD-378 210-         038273  0105  TITLE VARIES. SERIES CONTINUES UNDER A
                                    0110  DIFFERENT TITLE.

 1    LMSD-380 111-         038274  0115  -9          25 FEB 61-3 MAR 61        NCN
                                    0120  -10         (NOT ISSUED)
                                    0125  -11         11 MAR 61-17 MAR 61       NCN
                                    0130  -12         (NOT ISSUED)
                                    0135  -13         (NOT RECEIVED)
                                    0140  -14         1 APR 61-7 APR 61         NCN
                                    0145  -15         8 APR 61-14 APR 61        NCN
                                    0150  -16         (NOT RECEIVED)
                                    0155  -17         (NOT RECEIVED)
                                    0160  -18         29 APR 61-5 MAY 61        NCN
                                    0165  -19         6 MAY 61-12 MAY 61        NCN
                                    0170  -20         13 MAY 61-19 MAY 61       NCN
                                    0175  -21         20 MAY 61-26 MAY 61       NCN
                                    0180  -22         27 MAY 61-2 JUNE 61       NCN
                                    0185  -23         3 JUNE 61-9 JUNE 61       NCN
                                    0190  -24         10 JUNE 61-16 JUNE 61     NCN
                                    0195  -25         (NOT RECEIVED)
                                    0200  -26         24 JUNE 61-30 JUNE 61     NCN
                                    0205  -27         1 JULY 61-7 JULY 61       NCN
                                    0210  -28         8 JULY 61-14 JULY 61      NCN
                                    0215  -29         15 JULY 61-21 JULY 61     NCN
                                    0220  -30         22 JULY 61-28 JULY 61     NCN
                                    0225  -31         29 JULY 61-4 AUG 61       NCN
                                    0230  -32         5 AUG 61-11 AUG 61        NCN
                                    0235  -33         12 AUG 61-18 AUG 61       NCN

 1    LMSD-380 111-         039279  0000  -1  31 DEC 60-6 JAN 61                NCN
                                    0005  -2   7 JAN 61-13 JAN 61               NCN
                                    0010  -3   (NOT ISSUED)
                                    0015  -4   (NOT ISSUED)
                                    0020  -5  28 JAN 61-3 FEB 61                NCN
                                    0025  -6   4 FEB 61-10 FEB 61               NCN
                                    0030  -7   (NOT ISSUED)
                                    0035  -8   (NOT ISSUED)
                                    0040  TITLE VARIES.  SERIES CONTINUES UNDER A
                                    0045  DIFFERENT TITLE.

 1    LMSD-423 000-         038827  0000  -1  24 APR 59                         C-E-49
                                    0005  -2  17 JUNE 59 (SUPERSEDES LMSD-423 000-1)
                                    0010                                        C-F-501
                                    0015  -3  24 SEPT 59 (SUPERSEDES LMSD-423 000-2)
                                    0020                                        C-E-99

 1    LMSD-436 000-         037959  0000  -6  (SEE LMSD 429 253)
                                    0005  -7  (SEE LMSD 445 213)
                                    0010  -8  (SEE LMSD 445 242)
```

Fig. 12. Open-entry list

Coding and Tabulating Machine Processing of Physical Signs in Toxicity Tests

A system utilizing relatively simple machine methods is described for processing physical signs occurring in subacute and chronic toxicity tests in dogs and rats. Signs are coded by organ system and specific sign within organ systems. These codes are then entered on IBM cards. Signs that are not included in the code notation are entered in natural language. At the termination of the experiment, signs are printed out by sign and dosage group or by animal and dosage group. The system has simplified the analysis of the experiments and aided in evaluating the dose relationship of signs.

R. W. REICHARDT, S. P. SHER, R. FORD, R. B. ANDERSON, and E. E. VOGIN †

Merck Institute for Therapeutic Research and
Merck Sharp & Dohme Research Laboratories
West Point, Pennsylvania

• Introduction

Daily observations of the clinical condition, physical appearance, and behavior of animals used in subacute and chronic toxicity studies of compounds are made over periods extending from a few weeks to many months. The recording of these observations results in the accumulation of a considerable amount of data which must be analyzed and summarized in a form suitable for presentation as a final report. Hand processing these data has been a formidable task, since the number of animals may vary from a few dogs and rats in a subacute experiment to as many as 32 dogs and 200, or more, rats in a chronic experiment. The present report describes a system utilizing relatively simple machine equipment to process the data so that the occurrence of physical signs in rats and dogs may be more effectively analyzed and evaluated.

• Code Notations

An alphanumeric code for physical signs has been devised so that organ systems are designated by a number, and specific signs, referable to that organ system, are designated by a letter. Thus, 6 refers to the gastrointestinal tract and 6F and 6G signify tarry stools and emesis, respectively. A second digit refers to the number of days per week a particular sign occurred; thus, 6F2 signifies that tarry stools occurred on 2 days. The codes are similar for rats and dogs, but more positions have been designated for dogs because a greater variety of signs can be discerned in this species. The complete code for dogs in shown in Fig. 1. Within each organ system a "Z" designation exists for signs that are not otherwise specified in the code. A separate registry of "Z" codes in natural language is maintained. If a particular "Z" designation becomes significant, it is assigned an alphanumeric code.

The animals are individually housed and observed at frequent intervals each day. The physical signs and behavior of each animal are recorded daily in an official laboratory book. Animals exhibiting no overt signs are designated as "appears normal"; and in those studies in which a drug is administered at specified intervals (once daily, twice daily, etc.), signs are recorded before and after drug administration. Signs are coded at weekly intervals and transmitted to data processing.

The laboratory sheets are designed for an ordered system. Therefore, these sheets are arranged sequentially by animal number and are maintained in that order throughout the study. Similarly, IBM cards are kept in the same order.

† Present address: Food & Drug Research Laboratories, Inc., Maspeth, New York.

1 - CNS

A - Fine Tremor
B - Coarse Tremor
+ C - Behavioral Changes
D - Catatonia
E - Increased Activity
F - Decreased Activity, Sedation, etc.
G - Convulsions
H - Ataxia
I - Loss of Righting Reflex
J - Analgesia
* Z - NOS

2 - Eye/Ear

A - Ptosis
B - Miosis
C - Mydriasis
D - Lacrimation
E - Relaxation of nictitating membrane.
F - Scleral Injection
G - Conjunctivitis
H - Loss of Hearing
* Z - NOS

3 - Skin

A - Sores
B - Masses
C - Piloerection
D - Loose Hair
E - Alopecia
F - Other Skin Changes
* Z - NOS

4 - Respiratory

A - Nasal Discharge
B - Cyanosis
C - Decreased Respiratory Rate
D - Increased Respiratory Rate
E - Dyspnea
F - Panting
G - Dry Nose and Gums
H - Foul Odor of Breath
I - Rales
* Z - NOS

5 - Cardiovascular/Renal

A - Hyperemia, Ears and Gums
B - Blanched Ears and Gums
C - Edema
D - Hematuria
* Z - NOS

6 - Gastrointestinal

A - Abdominal Distension
B - Abdominal Tenderness
C - Soft Stool
D - Diarrhea
E - Frank Blood in Stools
F - Tarry Stools
G - Emesis
H - Emesis - Blood
I - Tenesmus
J - Ptyalism
K - Frequent Swallowing
L - Anorexia
* Z - NOS

7 - Reproductive

A - Increased Mammary Size
B - Lactation
C - Increased External Genitalia
D - Decreased External Genitalia
E - Priapism
F - Hyperemia, External Genitalia
G - Vaginal Discharge - Clear
H - Vaginal Discharge - Bloody
I - Estrus
* Z - NOS

8 - Musculo-Skeletal

A - Muscle Spasm
B - Muscle Tone Increased
C - Muscle Tone Decreased
D - Prostration
* Z - NOS

0 - General States

A - Poor Physical Condition
B - Interim Sacrifice
C - Found Dead
D - Sacrificed Moribund
E - Death Accidental
F - Escaped
* Z - NOS

+ = Write in sign.

* = Not otherwise specified.

FIG. 1. Dog sign code

• Card Layout and Program

Certain identifying information is included on each IBM card:

1. *TT #.* (Columns 3–8) Identifies the particular toxicity test, e.g., 65-0086 is the eighty-sixth subacute or chronic study conducted during 1965. Numbers are assigned sequentially.
2. *Dose Code.* (Columns 10–11) A two-digit designation indicates the dosage level that permits sequencing of cards according to dosage; e.g.:

00 = control
10 = low dose
20 = middle dose
30 = high dose

If a new dosage level is added, it may be coded in the proper dosage relation; 15 = dosage level between the low and middle dosage groups. The actual dosage utilized is recorded in the protocol for that experiment.

3. *Animal Number and Sex.* (Columns 13, 17, and 19)
4. *Sign Identity.* (Columns 68 and 69) A number and letter that designate the organ system and specific sign.
5. *Set Identification.* Each card contains a specified number of weeks of data. In rat studies, a set covers 9 weeks and in dogs, 6 weeks.

• Program

See Fig. 2.

Observations are recorded daily on an official laboratory sheet in natural language. At weekly intervals, the observations are coded, and a Xerox copy of the coded information is forwarded to data processing. Punched cards are prepared using an 026 Printing Punch. One card is prepared for each sign. The sign is punched into the appropriate field for the current week. The cards are then sequenced: first by dosage group, next by animal number, and finally by sign.

The deck of newly prepared data is then collated against the file deck using a routine of "merge if matched." Thus, cards are delivered into one of three pockets:

1. Nonmatching cards representing the first appearance of the sign in a given animal.
2. Nonmatching cards representing the absence of a previously observed sign.
3. Matching cards representing signs that have been previously observed. The new card is filed in front of the earlier observation.

The matching cards are posted using an "alternate program routine" with an 026 Printing Punch. The weekly card is passed to the reading station, and the weekly sign is automatically copied into the appropriate field for the current week. After posting, the duplicate cards are removed from the deck using a 101 Statistical Machine (or a collator) and then discarded. The newly posted cards are merged with the nonmatching cards (1 and 2 above), and the file containing the new data is then back in original sequence.

• Use of Sign Data in Analysis of Drug Results

At the conclusion of the toxicity test or at an interim period, a printout of all signs is furnished. Two types of printouts are prepared:

1. Sequenced by sign code and dosage group, sex, and animal number respectively.
2. Sequenced by dosage group, sex, animal number, and sign respectively.

The use of the first printout (grouping of signs by dosage group), permits an evaluation of the relationship of sign frequency to dosage group. Thus, it is possible to quickly detect signs scattered through both control and drug-treated groups which are probably unrelated to treatment. In addition, this printout permits the investigator to detect any apparent dose relationship in the occurrence of any particular sign. In Fig. 3, the order of occurrence of sign 6L (anorexia) is dose code 30 dose code 20 dose code 10 dose code 40 dose code 00. The onset and duration of any given signs can be easily determined.

The second printout, (Fig. 4), which lists all signs occurring in a given animal, permits the investigator to evaluate the physical condition and behavior of that animal. Furthermore, it is possible to follow either the progression or regression of the severity of treatment within a given animal and to note the possible relationship between signs observed.

In addition to the sign code we are also entering rat body weight, dog body weight, food consumption, water intake, and urine output on IBM cards. Dog water, urine, and food data are entered daily, and weekly summaries are prepared. Since body weights are recorded once weekly, the single entry is, in effect, a weekly summary. The use of machine methods for this data has substantially reduced the amount of hand copying, as well as typing of final tables. Plans are under way to include machine processing of pathologic data.

Related Literature

DIETRICH, E. V., Machine Retrieval of Pharmacological Data, *Science*, 132:1556–1557 (1960).

DIETRICH, E. V., Data Analysis and Punched Cards, *Toxicology and Applied Pharmacology*, 8:338 (1966).

MCKELVIE, D. H., and F. T. SHULTZ, Methods of Observing and Recording Data in Long-term Studies on Beagles, *Laboratory Animal Care*, 14:118–124 (1964).

OWEN, G., and H. P. K. AGERSBORG, JR., Data-processing Techniques in Toxicology, *Toxicology and Applied Pharmacology*, 8:349 (1966).

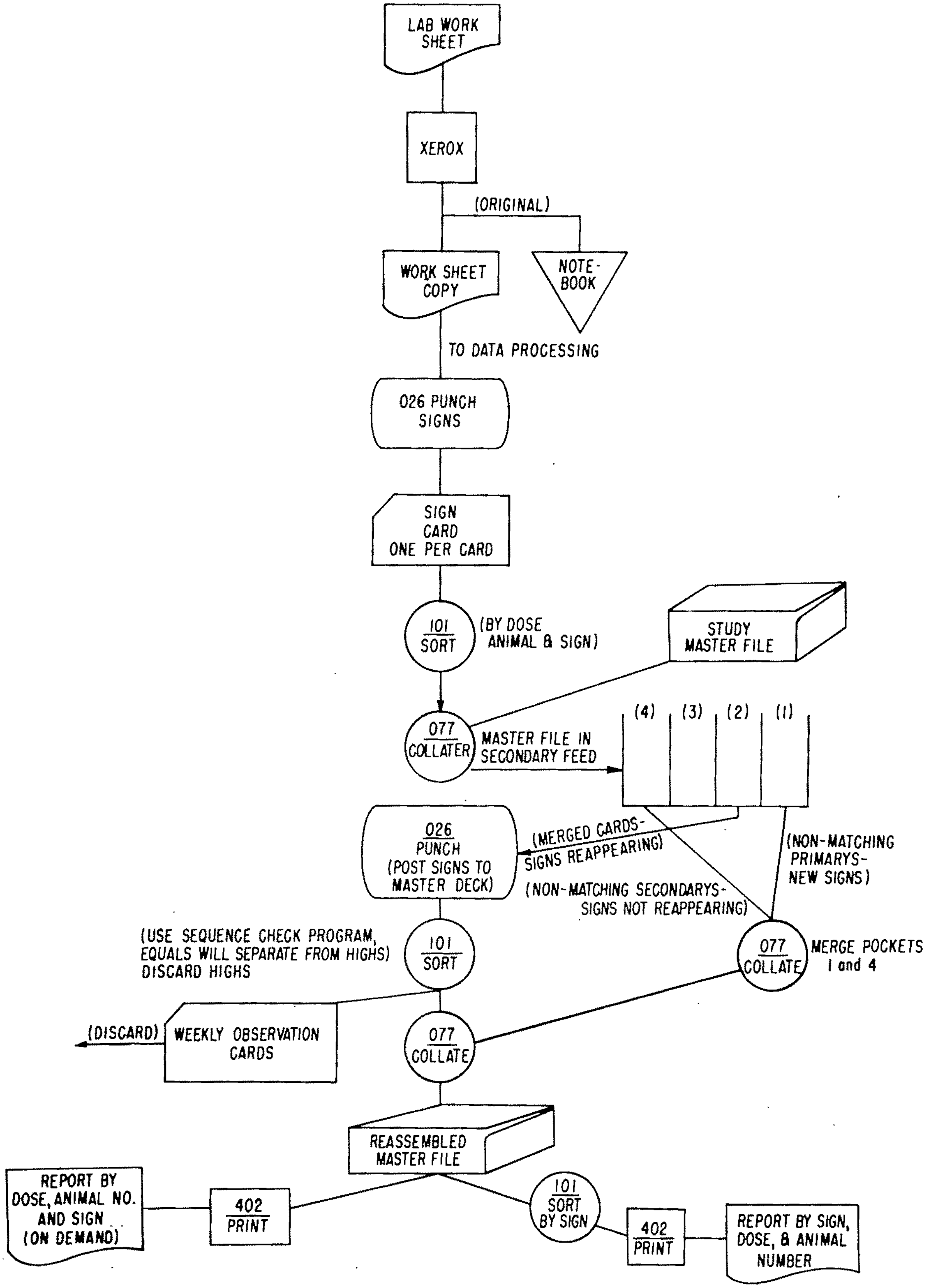
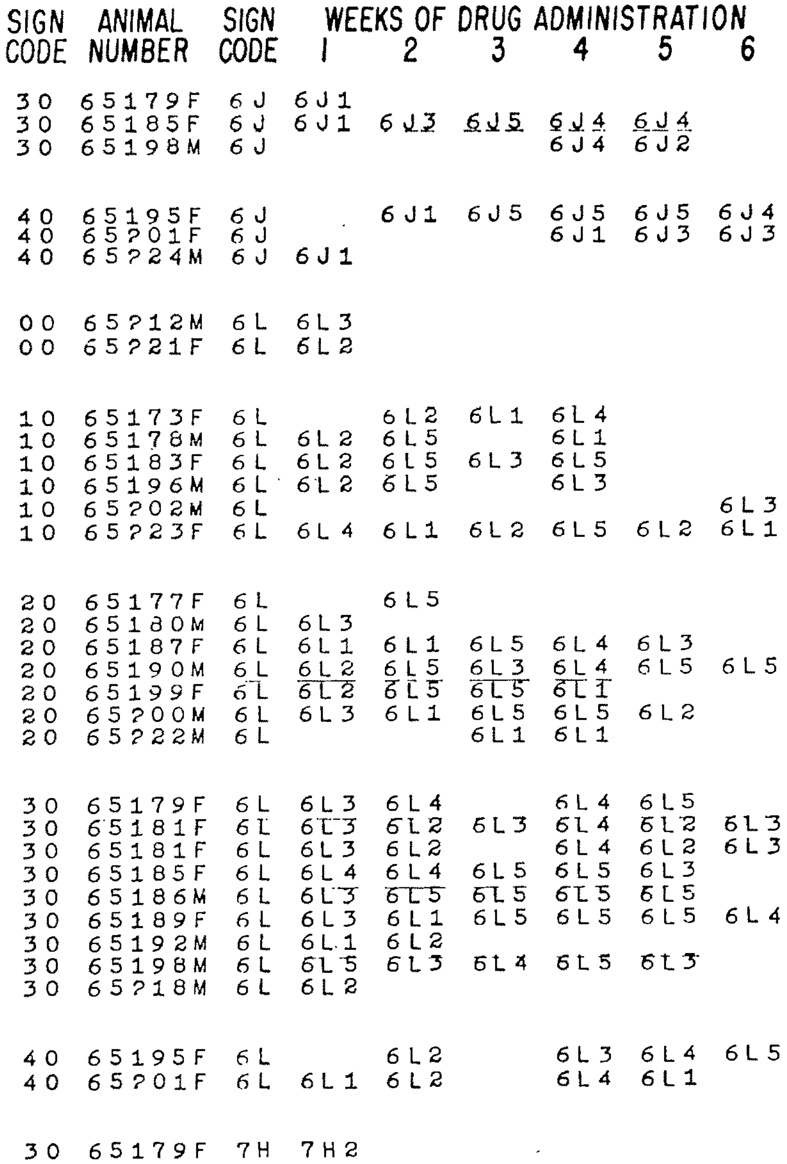

LAB WORK SHEET
XEROX
(ORIGINAL)
WORK SHEET COPY
NOTE-BOOK
TO DATA PROCESSING
026 PUNCH SIGNS
SIGN CARD ONE PER CARD
101 SORT
(BY DOSE ANIMAL & SIGN)
STUDY MASTER FILE
077 COLLATER
MASTER FILE IN SECONDARY FEED
(4)
(3)
(2)
(1)
026 PUNCH (POST SIGNS TO MASTER DECK)
(MERGED CARDS-SIGNS REAPPEARING)
(NON-MATCHING SECONDARYS-SIGNS NOT REAPPEARING)
(NON-MATCHING PRIMARYS-NEW SIGNS)
(USE SEQUENCE CHECK PROGRAM, EQUALS WILL SEPARATE FROM HIGHS) DISCARD HIGHS
101 SORT
077 COLLATE
MERGE POCKETS 1 and 4
(DISCARD)
WEEKLY OBSERVATION CARDS
077 COLLATE
REASSEMBLED MASTER FILE
REPORT BY DOSE, ANIMAL NO. AND SIGN (ON DEMAND)
402 PRINT
101 SORT BY SIGN
402 PRINT
REPORT BY SIGN, DOSE, & ANIMAL NUMBER

FIG. 2. Flow diagram of machine processing of sign data in toxicity tests



SIGN CODE	ANIMAL NUMBER	SIGN CODE	WEEKS OF DRUG ADMINISTRATION 1	2	3	4	5	6
30	65179F	6J	6J1					
30	65185F	6J	6J1	6J3	6J5	6J4	6J4	
30	65198M	6J				6J4	6J2	
40	65195F	6J		6J1	6J5	6J5	6J5	6J4
40	65?01F	6J				6J1	6J3	6J3
40	65?24M	6J	6J1					
00	65?12M	6L	6L3					
00	65?21F	6L	6L2					
10	65173F	6L		6L2	6L1	6L4		
10	65178M	6L	6L2	6L5		6L1		
10	65183F	6L	6L2	6L5	6L3	6L5		
10	65196M	6L	6L2	6L5		6L3		
10	65?02M	6L						6L3
10	65?23F	6L	6L4	6L1	6L2	6L5	6L2	6L1
20	65177F	6L		6L5				
20	65180M	6L	6L3					
20	65187F	6L	6L1	6L1	6L5	6L4	6L3	
20	65190M	6L	6L2	6L5	6L3	6L4	6L5	6L5
20	65199F	6L	6L2	6L5	6L5	6L1		
20	65?00M	6L	6L3	6L1	6L5	6L5	6L2	
20	65?22M	6L			6L1	6L1		
30	65179F	6L	6L3	6L4		6L4	6L5	
30	65181F	6L	6L3	6L2	6L3	6L4	6L2	6L3
30	65181F	6L	6L3	6L2		6L4	6L2	6L3
30	65185F	6L	6L4	6L4	6L5	6L5	6L3	
30	65186M	6L	6L3	6L5	6L5	6L5	6L5	
30	65189F	6L	6L3	6L1	6L5	6L5	6L5	6L4
30	65192M	6L	6L1	6L2				
30	65198M	6L	6L5	6L3	6L4	6L5	6L3	
30	65?18M	6L	6L2					
40	65195F	6L		6L2		6L3	6L4	6L5
40	65?01F	6L	6L1	6L2		6L4	6L1	
30	65179F	7H	7H2					

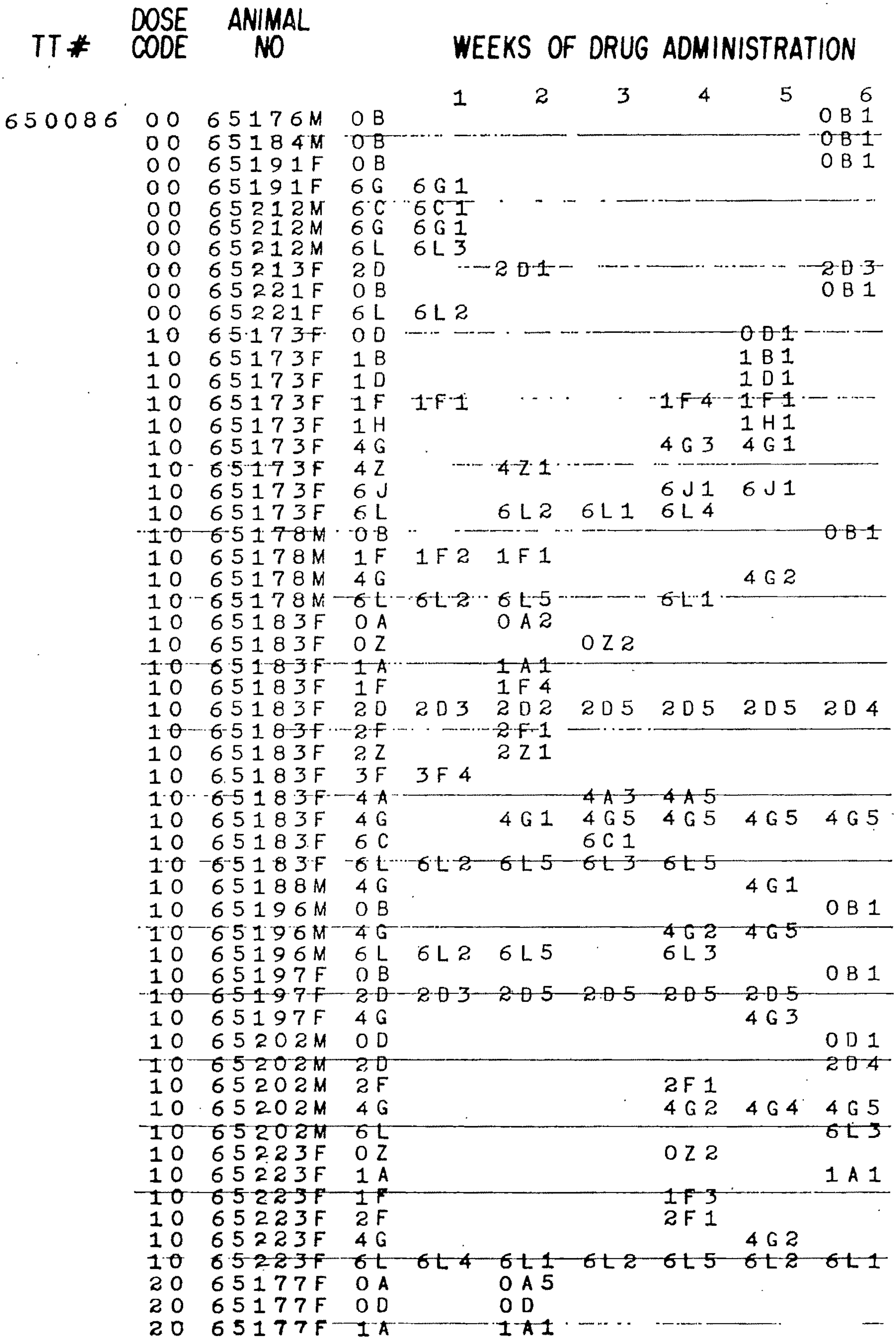
FIG. 3. Physical signs in dogs listed by signs and dosage group

RUSSELL, T. J., W. G. WAGGONER, and E. B. GASSER, Application of Automatic Data Acquisition and Digital Computation to the Reduction of Data Generated During Techniques in Toxicology, *Toxicology and Applied Pharmacology,* 8:349 (1966).

SMALL, R. M., and R. C. ANDERSON, Semi-automatic Recording and Electronic Processing of Chronic Rat Toxicity Data, *Laboratory Animal Care,* 15:345–353 (1965).

SMITH, J. O., and J. MELTON, Autopsy Diagnoses by Computer Technique, *Journal of the American Medical Association,* 188:953–962 (1964).

WAGGONER, W. G., T. J. RUSSELL, and E. B. GASSER, Data Acquisition and Processing System for Animal Toxicology, *Federation Proceedings,* 25:447 (1966).



TT #	DOSE CODE	ANIMAL NO		WEEKS OF DRUG ADMINISTRATION					
				1	2	3	4	5	6
650086	00	65176M	0B						0B1
	00	65184M	~~0B~~						~~0B1~~
	00	65191F	0B						0B1
	00	65191F	6G	6G1					
	00	65212M	6C	~~6C1~~					
	00	65212M	6G	6G1					
	00	65212M	6L	6L3					
	00	65213F	2D		~~2D1~~				~~2D3~~
	00	65221F	0B						0B1
	00	65221F	6L	6L2					
	10	65173F	0D					~~0D1~~	
	10	65173F	1B					1B1	
	10	65173F	1D					1D1	
	10	65173F	1F	~~1F1~~			~~1F4~~	~~1F1~~	
	10	65173F	1H					1H1	
	10	65173F	4G				4G3	4G1	
	10	65173F	4Z		~~4Z1~~				
	10	65173F	6J				6J1	6J1	
	10	65173F	6L		6L2	6L1	6L4		
	~~10~~	~~65178M~~	~~0B~~						~~0B1~~
	10	65178M	1F	1F2	1F1				
	10	65178M	4G					4G2	
	10	65178M	~~6L~~	~~6L2~~	~~6L5~~		~~6L1~~		
	10	65183F	0A		0A2				
	10	65183F	0Z			0Z2			
	~~10~~	~~65183F~~	~~1A~~		~~1A1~~				
	10	65183F	1F		1F4				
	10	65183F	2D	2D3	2D2	2D5	2D5	2D5	2D4
	~~10~~	~~65183F~~	~~2F~~		~~2F1~~				
	10	65183F	2Z		2Z1				
	10	65183F	3F	3F4					
	10	~~65183F~~	~~4A~~			~~4A3~~	~~4A5~~		
	10	65183F	4G		4G1	4G5	4G5	4G5	4G5
	10	65183F	6C			6C1			
	10	~~65183F~~	~~6L~~	~~6L2~~	~~6L5~~	~~6L3~~	~~6L5~~		
	10	65188M	4G					4G1	
	10	65196M	0B						0B1
	~~10~~	~~65196M~~	~~4G~~				~~4G2~~	~~4G5~~	
	10	65196M	6L	6L2	6L5		6L3		
	10	65197F	0B						0B1
	~~10~~	~~65197F~~	~~2D~~	~~2D3~~	~~2D5~~	~~2D5~~	~~2D5~~	~~2D5~~	
	10	65197F	4G					4G3	
	10	65202M	0D						0D1
	~~10~~	~~65202M~~	~~2D~~						~~2D4~~
	10	65202M	2F				2F1		
	10	65202M	4G				4G2	4G4	4G5
	~~10~~	~~65202M~~	~~6L~~						~~6L3~~
	10	65223F	0Z				0Z2		
	10	65223F	1A						1A1
	~~10~~	~~65223F~~	~~1F~~				~~1F3~~		
	10	65223F	2F				2F1		
	10	65223F	4G					4G2	
	~~10~~	~~65223F~~	~~6L~~	~~6L4~~	~~6L1~~	~~6L2~~	~~6L5~~	~~6L2~~	~~6L1~~
	20	65177F	0A		0A5				
	20	65177F	0D		0D				
	20	65177F	~~1A~~		~~1A1~~				

Fig. 4. Physical signs in dogs listed by dosage group and animal number

Subject Searching with *Science Citation Index*: Preparation of a Drug Bibliography Using *Chemical Abstracts, Index Medicus,* and *Science Citation Index* 1961 and 1964*

A bibliography on the drug thalidomide was prepared through a search of *Chemical Abstracts* (CA) and *Index Medicus* (*IM*) for the years 1956–1964, which took 14.6 hours. This was compared with a similar bibliography prepared through a search of *Science Citation Index* (*SCI*) 1961 and 1964, carried out for an equal length of time according to search procedures described, in an effort to determine if and how *SCI* might be helpful in subject searches. A satisfactory procedure for manual search of *SCI* was developed. The cumulative number of references found through *SCI* plotted against time gave a convex curve, while corresponding data from the conventional indexes gave a linear response. For periods up to eight hours, *SCI* yielded more references than did either *IM* or CA. However, at 14.6 hours *SCI* did not produce all references obtainable through CA-*IM*. Each of the three indexes produced a high percentage of unique references, *SCI*'s being the highest. *SCI*-1964's coverage of articles published in 1964 was more complete than was *IM*-1964's. CA-*IM* gave superior coverage of chemical papers, patents, and papers in the less common languages. *SCI*'s coverage of pharmacological papers was superior to that of either CA or *IM*. In this search *SCI* and conventional indexes could be profitably used together; each produced a large number of references not to be found in the other. In this search *SCI* was not appreciably less efficient in retrieving drug references than were CA and *IM*; for short time intervals, it was more efficient. More general conclusions must await further investigation.

CAROL C. SPENCER

Institute for Advancement of Medical Communication
Philadelphia, Pennsylvania

• Introduction

The work reported here is the result of a proposal to compare use characteristics, i.e., the time needed to search and the result of search, of

Science Citation Index (*SCI*) 1961 and 1964
with
Chemical Abstracts (*CA*) and *Index Medicus* (*IM*)

for preparing a bibliography on a particular drug. Pharmaceuticals was chosen as the field to be represented because it was considered a good example of a multidisciplinary field. Thalidomide was chosen as the topic drug because its properties were well publicized, and appearance of references to it were timely from the standpoint of coverage by *SCI*. The time period chosen for searching was from 1956 (the date of the first publication concerning the drug) through 1964 (the last year for which complete indexes for all three services were available at the outset of this project). The experiment was designed so that time to be spent searching *SCI* was determined by the time required to complete a thorough *CA-IM* conventional subject search for references to thalidomide. By entering the conventional indexes (*CA* and *IM*) with a chemical or generic name, one could expect optimum results from them; that is, there were no problems of ambiguity either in designating a name for what was wanted or in determining how it would have been indexed.

Some of the questions explored are: (1) How does searching *SCI* compare (in efficiency and output volume) with searching *CA-IM* for the same topic? (2) Can *SCI* produce references not found by *CA* or *IM*? (3) Is there

* This paper is based on an M.S. thesis submitted to Drexel Institute of Technology, October 1966.

any appreciable difference in the outputs? If so, what are the reasons? (4) What is the nature of any unique output? (5) Can *SCI* and conventional indexes be combined in some way to produce results superior to either type alone?

Related Work

Garfield (*1*) has suggested using citation indexes for subject searching, but some think that the value of using them this way is yet to be proven. Martyn (*2*) states without qualification that, as a retrieval tool, *Science Citation Index* (*3*) is "not as efficient" as the more conventional indexes, however well it may function as an access tool.

Touloukian (4) argues for citations, but not necessarily for citation indexes, when he suggests using abstracting journals to locate recent papers, then using their bibliographies to trace additional papers. He has found this to be more efficient than searching the abstracting journals throughout the entire time period.

Waldhart (*5*), who compiled a bibliography on lasers using seven conventional indexes and *SCI*-1961, found that in both *SCI* and conventional indexes he could find references at the rate of three minutes per reference. He did not use the productive technique of "cycling" (using the bibliography of the starting reference to provide additional access points) but began with a single reference as his entry into *SCI*-1961 and collected references that cited it. Cycling might have lowered his production time below three minutes per reference but would have required access to the original journals.

Garfield, Sher, and Torpie (*6*) used the bibliographies of their primary references on DNA in order to complete the map of the citation network, but their concern was with the relationship between articles, rather than the accumulation of a large number of additional references in a limited time.

Baker (*7*) used *Index Chemicus* as a source of primary references on alkaloids and then obtained additional references by feeding the primary references into *SCI*-1964. She did not use the cycling technique, and her concern was with the nature of the output rather than with the time required to produce it.

General Considerations

The main difficulties in using *SCI* as a subject index appear to be:

1. *Unfamiliar format.* Access to *SCI* is gained by means of a starting reference, i.e., a reference the searcher already knows, rather than by means of a subject heading, as in conventional indexes. The starting reference must have the author's name spelled correctly in most instances and may require at least his initials to distinguish his works from those of other authors with the same surname; in *SCI* authors are listed alphabetically. Following each *cited* author is a list of his papers. After each of his papers is a list of other papers that cite it. In another section, the Source Index, the *citing* authors are listed alphabetically, together with the citations and titles of their papers, the type of article (review, editorial, etc.), and the number of references that the particular paper cites. The format is well illustrated and explained in the Institute for Scientific Information's training publication (*8*). After a bit of practice with this manual and with the *SCI* itself, the unfamiliar format was no problem.

2. *No direct subject approach.* The lack of subject approach can be compensated for by consulting appropriate conventional indexes to obtain early landmark papers or recent reviews, if the searcher does not know the name of an author in the field.

3. *Noise, irrelevant references.* A hypothetical starting reference A may be cited, for example, by papers 1–10. Perhaps only 2 of these 10 actually deal with the aspect of paper A in which the searcher is interested; the others may refer to another aspect, or to a technique developed by the author of A which the citing author has applied in another subject area. If this should happen repeatedly, the searcher may accumulate a large number of references irrelevant for *his* purposes. Careful selection of a homogeneous starting reference (one which deals with only one subject) can minimize this problem. The Source Index to *SCI* will contain the full titles of papers 1–10 and will thereby give some indication of their probable relevance. Garfield (*1*) suggests that if noise is a problem with *SCI*, it is possible for the searcher to record only references that have cited *two or more* starting references, thus increasing the chance of relevance. In planning a procedure for using *SCI* most efficiently, one must block off or defer the questionably relevant leads (such as papers whose titles do not indicate positive relevance) at least until the more productive approaches (such as papers whose titles do indicate positive relevance) have been exhausted.

4. *Excessive time consumption.* The average working literature searcher or reference librarian must be very much aware of time; for routine searches he is unlikely to use an index that is excessively time consuming, however great its other advantages. Excessive search time may be a result of the noise factor and also of suboptimum search strategy. The various steps or operations within the search process must be correctly evaluated for their relative productivity and priorities for their use assigned accordingly.

The advantages of using *SCI* for subject searching would seem to be:

1. *No terminology problem.* If the lack of subject approach is a disadvantage, it also has favorable aspects. There is no need to guess how an indexer might have indexed the desired material if one knows the author and other (even fragmentary) information about one or more papers dealing with it.

2. *Interdisciplinary coverage.* The overlapping areas

between classic disciplines are not reliably covered by the conventional indexes; therefore *SCI*'s complete coverage of the interdisciplinary journals such as *Nature* and *Science* is a valuable asset.

3. *Complete coverage of "covered" journals.* Conventional indexes are necessarily selective in their coverage of many journals, and their criteria for selection and omission are by no means always obvious. Letters to the editor and convention proceedings are almost universally slighted. *SCI* indexes every item in journals it covers, even errata notices.

• Methods

Conventional Search

Chemical Abstracts (*CA*) was searched on the heading "phthalimide, N-(2,6-dioxo-3-piperidyl)" from 1956 (the date of the first publication on thalidomide) through 1964. This yielded 110 references in 5.6 hours.

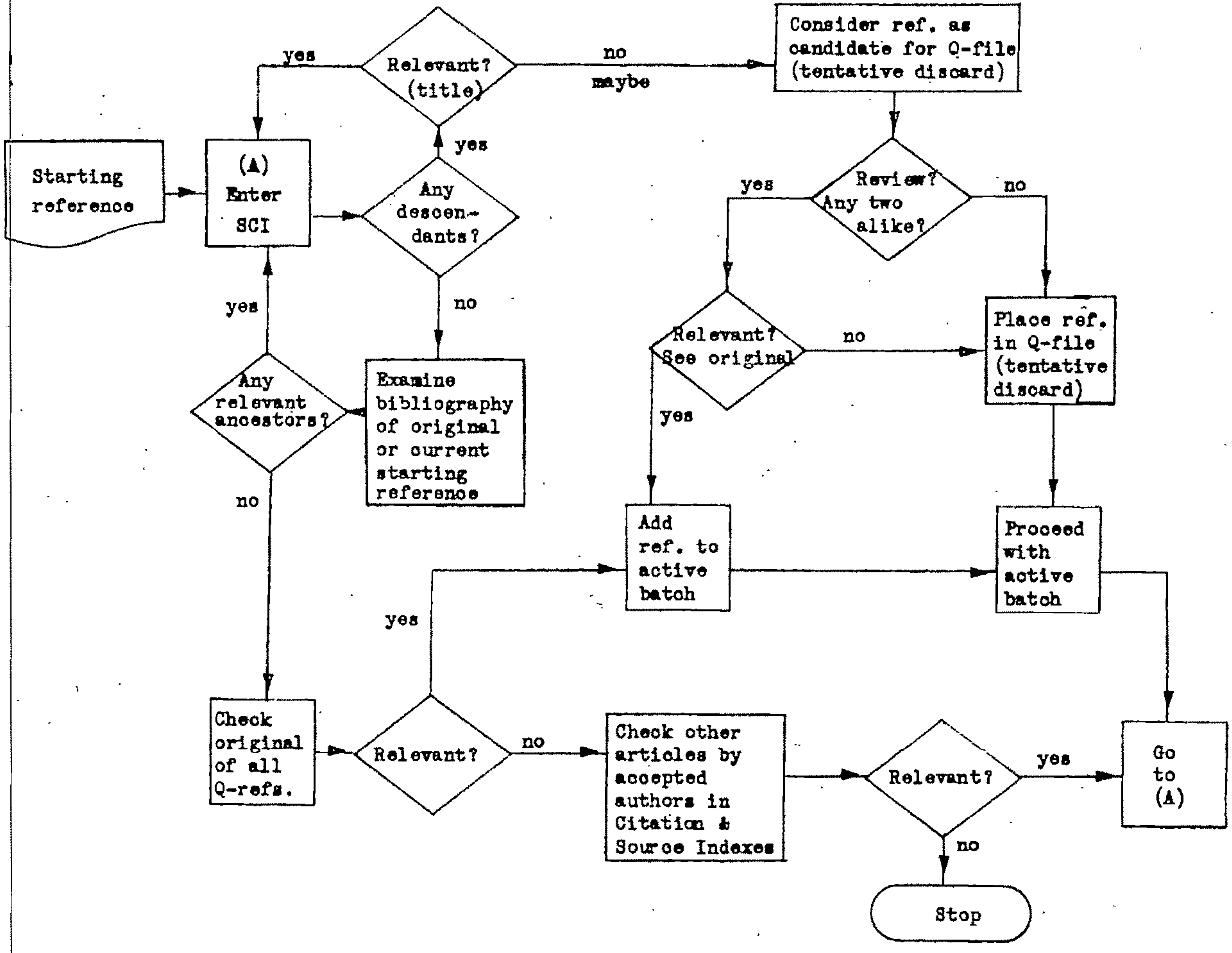
Index Medicus (*IM*) was searched on the heading "thalidomide" from 1963 through 1964, producing 275 references in 6.5 hours. Since the heading "thalidomide" was not used before 1963, *IM* was also searched on the heading "hypnotics and sedatives" for the years 1956 through 1962, and titles were scanned for the word "thalidomide" or its synonyms. (The probable incompleteness of this latter step may be estimated from the fact that of the 110 references from *CA* and the 275 from *IM*-1963–64, 70% could be judged relevant from the titles alone.) The search of *IM* produced a total of 370 references in 9 hours.

Search Time "T" Hours

Since the time required to search both *CA* and *IM* from 1956–1964 (resulting in 429 different references) was 14.6 hours, according to the design of the experiment, that amount of time was taken as a unit for measuring the time used for the *SCI* search: $T=14.6$ hours. (In all cases, search time included copying time.)

Search of "Science Citation Index" (*SCI*)

See Figure 1 for details about the basic procedure used.



Relevant? (title)
yes
no
maybe
Consider ref. as candidate for Q-file (tentative discard)
Starting reference
(A) Enter SCI
Any descendants?
yes
no
Review? Any two alike?
yes
no
yes
Any relevant ancestors?
Examine bibliography of original or current starting reference
Relevant? See original
no
yes
Place ref. in Q-file (tentative discard)
no
Add ref. to active batch
Proceed with active batch
yes
Check original of all Q-refs.
Relevant?
no
Check other articles by accepted authors in Citation & Source Indexes
Relevant?
yes
no
Go to (A)
Stop

Fig. 1. Flow diagram for *SCI* search procedure

Criterion for relevance: In order to avoid subjective judgments, an article was considered relevant if its full text contained *any* mention of thalidomide or its synonyms.

First Run (hereafter designated as *SCI*-1)

Chosen as the starting reference was the first significant publication on thalidomide listed in *CA:*

> Kunz, W., N-phthalylglutamic acid imide: Experimental studies on a new synthetic product with sedative properties, *Arzneimittelforschung* 6 (8):426–430 (1956)

(Other relevant references were on hand in case this reference failed to be cited.) This starting reference was used as access into *SCI* 1961 and 1964. (*SCI's* for 1962 and 1963 have not been published.)

Step 1. All papers that cited this starting reference were recorded. The references (author, journal, page, and year) were looked up in the Source Index of *SCI* to locate titles for the articles, to assess their possible relevance. When positive relevance was determined from the title, each relevant descendant (paper that cited the starting reference) became a candidate for access into *SCI* as a *new* starting reference, and so on, until no more references resulted. If positive relevance could not be determined from the title, the reference was placed in a tentative discard (Q-file) *unless* the reference was a review article (as indicated in the Source Index) *or* had appeared for the second time. If a reference was:

1. *A review article,* it was not put in tentative discard, even if the title did not indicate relevance. All reviews were immediately consulted in the original to determine relevance.
2. *A candidate for tentative discard for the second time,* that is, if it was a descendant of *two* relevant articles and therefore much more likely to be relevant than a paper descended from only one relevant reference, its original was immediately consulted.

Members of these two classes of articles, if relevant, were immediately subjected to Steps 1 and 2. Review articles were unusually productive of additional references and were processed out of the usual order so as to obtain as many references as quickly as possible. At the point when no more references could be obtained from these papers citing the starting reference, the accumulation contained the original starting reference and a number of its descendants.

Step 2. All of the relevant references so far accumulated by searching *SCI* were consulted in the journal in which they originally appeared. Their bibliographies were scanned, and all ancestor references (papers which these papers cited) which appeared relevant (either from their titles or from statements about them in the text of the article) were selected and recorded. These ancestor references were also candidates for access to *SCI;* their use as new starting references begins the process of "cycling." Step 1 was repeated on the references located through Step 2 until no more references could be obtained.

Step 3. When no more references could be obtained from iterating through Steps 1 and 2, those references that had accumulated in the tentative discard file were checked in the journal in which they originally appeared to determine their relevance. If relevant, they were also used to access *SCI* as above.

Step 4. If time permitted, the most productive authors were looked up in both the Citation Index and the Source Index in hope of finding more relevant references.

All this was done for 14.6 hours, time equal to that spent on the conventional-index search. Elapsed time and the number of relevant references obtained were recorded at the end of each work period.

Additional Runs

In order to try other ways of using *SCI,* a second and third run were planned in order to see if a larger output could be obtained by using inputs other than one starting reference, while using the same basic procedure.

Second Run (hereafter designated as *SCI*-2)

This was the same as the First Run except for the initial input. For *SCI*-2 the input was the review articles obtained in a quick search of *CA*-1962–64 (for reviews only, as classified by *CA*) and their bibliographies.

Third Run (hereafter designated as *SCI*-3)

This was the same as the First Run except for the initial input. For *SCI*-3 the input was the entire product of the *CA-IM* search, 429 references. They were processed in alphabetical order. Descendants of these starting references (from Steps 1 and 2) were checked first against the list of 429 *CA-IM* articles (List A), a quick, if incomplete way to determine relevance. If necessary, their titles were then checked in the Source Index, as in the First Run. In all other respects the procedure followed was the same as in the First Run.

Compilation of Composite File From All Sources, CA, IM, SCI-1, SCI-2, SCI-3

Reference cards from all five sources were pooled, and a composite file containing one card per reference was made on marginally-punched tabulating cards (*9*). On each card was recorded, coded, and punched each of the five sources through which the reference was located.

For those articles not found through *CA* or *IM,* the indexes were rechecked to determine *why* they were not found:

O—Indexed under other heading. (Found in author index.)

J—Journal not covered.
L—Indexed later, after 1964.
S—"Selective coverage." (Journal listed as covered, but article not found in author index in year of publication nor two following years.)

This information was recorded, coded, and punched into the tabulating card.

Each article was classified as to whether it was primarily chemical, clinical, or pharmacological. It was also classified as to whether it was:

1. Case history or original study
2. Review (as classified by *SCI, CA*, or *IM*)
3. Patent
4. Editorial
5. Miscellaneous

The language and country of origin was recorded, coded, and punched in the tabulating card for all articles. In *IM*, the original language is explicitly stated, if other than English. In *CA*, the original language is explicitly stated if (1) it is other than that which would be inferred from the country of origin of the journal, or (2) it could not be so inferred, as with Swiss journals. For articles from *SCI* runs, the original journal was consulted.

All references from *SCI* runs which could not be verified or which proved incorrect were discarded. Twenty-one items were so discarded, yielding a total composite file of 632 articles.

• Results

Efficiency

Figure 2 shows the hourly outputs of the *CA, IM, SCI*-1 and *SCI*-2 searches, and Fig. 3 shows the cumulative number of references vs. search time for the same four searches. (*SCI*-3, a supplemental exhaustive search, is not included in the efficiency comparison since a time limit is not appropriate to this type search.)

It can be seen from Figs. 2 and 3 that if one chose to spend 8 hours or less on a thalidomide search, *SCI*-2 would yield the highest number of references in that time, and that most of those references could be obtained in the first 4 hours.

SCI-1 proved superior in efficiency to the conventional indexes for up to 6 hours, most of the references being obtained in the first 3 hours.

Table 1 shows the average time required to obtain one reference by each procedure. From this it appears that *SCI*-1 and *SCI*-2 were not appreciably less efficient than conventional indexes for a search on thalidomide; for the short search time intervals most often encountered in practical situations, they were more efficient.

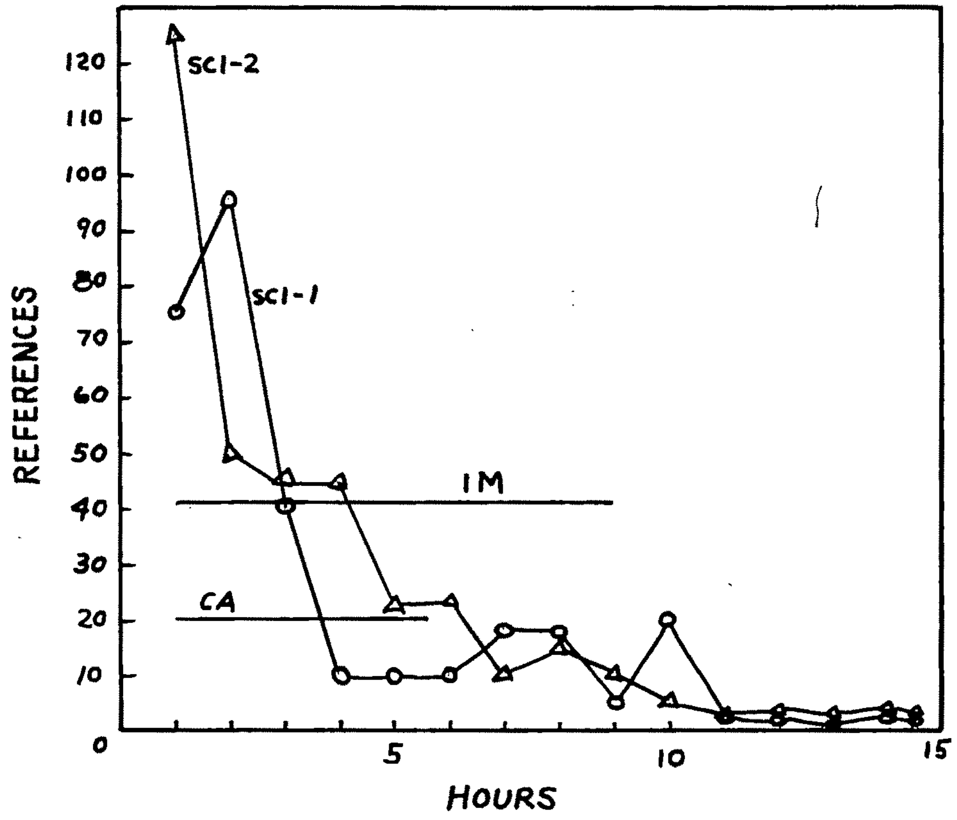


FIG. 2. Hourly output in references for *CA, IM, SCI*-1, and *SCI*-2 searches

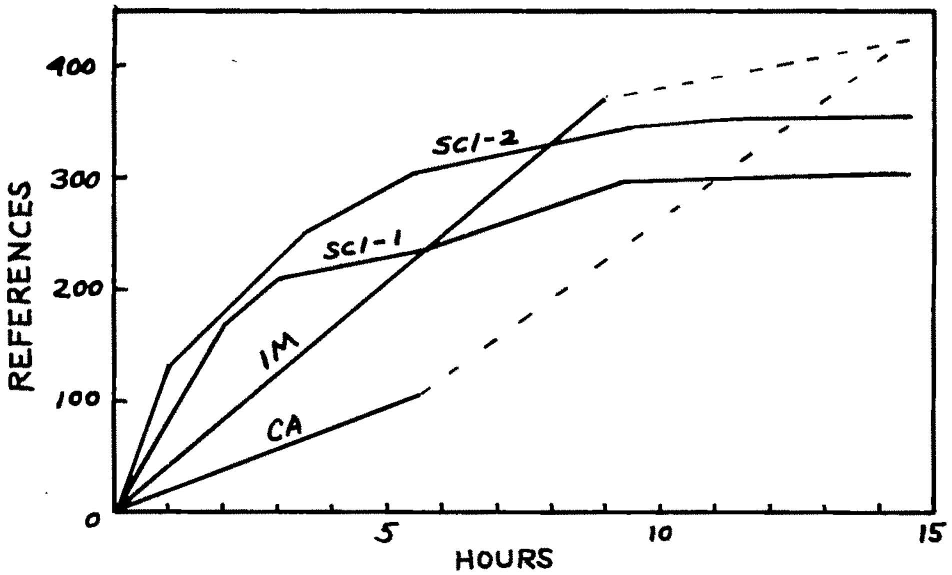


FIG. 3. Cumulative number of references vs. search time for *CA, IM, SCI*-1, and *SCI*-2 searches. Dotted lines indicate combined product of *CA* and *IM* searches.

SCI-2 was more efficient than *SCI*-1 for locating references on thalidomide for the period of time covered in this experiment; in *SCI*-2, the review articles are processed early and review articles are the richest source of new references. This effect is illustrated in Table 2, which shows which part of the search procedure produced the references in *SCI*-1 and *SCI*-2. The great productivity of the relevant review articles in *SCI*-2 emphasizes the value of the cycling technique (which requires examination of the original article) and confirms the value of preferential treatment for *possibly* relevant review articles. The cycling technique adds to efficiency, but requires access to original journals, impossible in some situations. (However, Table 4 will indicate that quite a number of articles could be found using only *SCI, Lancet,* and *British Medical Journal.*)

TABLE 1. Average time to obtain one reference

Hours	Minutes/Reference (average) *CA*	*IM*	*SCI*-1	*SCI*-2	*IM-CA*
1	3.1	1.5	0.80	0.48	
2	3.1	1.5	0.71	0.69	
3	3.1	1.5	0.86	0.82	
4	3.1	1.5	1.1	0.91	
5	3.1	1.5	1.3	1.04	
6		1.5	1.5	1.17	
7		1.5	1.63	1.32	
8		1.5	1.75	1.44	
9		1.5	1.9	1.6	
10			2.0	1.72	
11			2.18	1.87	
12			2.37	2.04	
13			2.55	2.20	
14			2.73	2.33	
14.6			2.85	2.43	2.03

RELATIVE COMPLETENESS

Table 3 and Figure 4 show the overlap among indexes and the large amount of material covered uniquely by each index.

Figure 5 indicates the relative completeness of the searches done if the composite total of 632 references is taken as 100%. This graph suggests that for an exhaustive search, a combination of *CA-IM* and *SCI*-2 would give the most complete results. However, *SCI*-3 (including *CA-IM*) might very well yield the most complete results if time were *not* limited. Ten references were obtained *only* through *SCI*-3, and eventually all

TABLE 2. Productivity of various steps in *SCI* search procedure

Point at which reference was called relevant	*SCI*-1		*SCI*-2	
	No.	%	No.	%
From relevant reviews:				
Relevance indicated in text of review article	97	32	269	77
From Source Index:				
Title contained word "thalidomide" or synonym	31	10	30	9
Original consulted:				
Source Index indicated paper was a review article	11	4	11	8
Cited twice by known relevant papers	88	29	5	1
Q-file: tentative discard	75	25	33	10

TABLE 3. Overlap among indexes and unique coverage for each index

Source	Number of references	Percent of 632	Percent unique references
CA only	48	7.5	
IM only	196	31.0	
SCI only	203	32.0	
(All exclusives)	(447)	(71.0)	
CA and *IM*	21	3.4	
CA and *SCI*	11	1.8	
IM and *SCI*	123	19.5	
CA, IM, and *SCI*	30	4.8	
Total	632	100.0	
Total *CA*	110		43.6
Total *IM*	370		53.0
Total *SCI*	367		55.3

references located by *SCI*-1 and *SCI*-2 would have been located by *SCI*-3, given enough time.

CHARACTER OF OUTPUT

Table 4 shows that the vast majority (84%) of the thalidomide references were case histories or original studies. If they were *not* indexed by the conventional indexes, the usual reason was "selective coverage" (84% for CA, 56% for *IM*).

In this sample of 632 items, chemical papers were well covered by *CA* (97% as compared to 13% for *IM* and

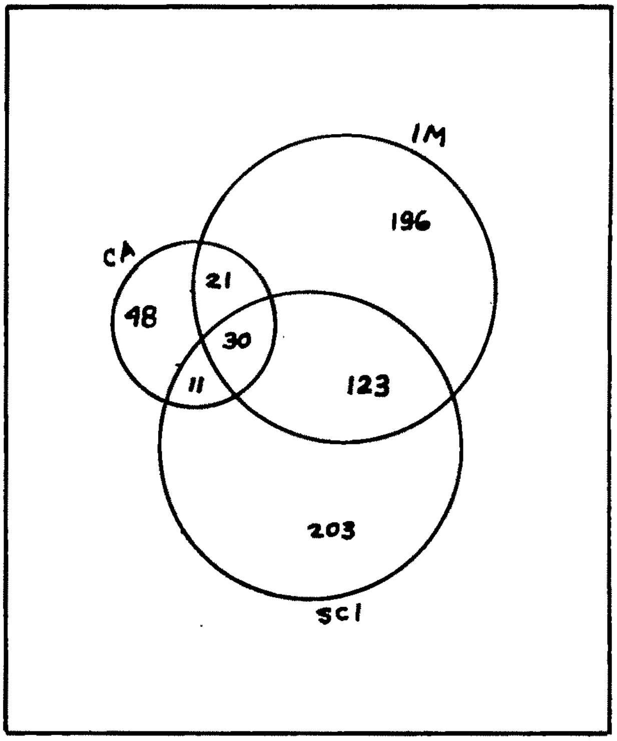


FIG. 4. Overlap among indexes and unique coverage of *CA, IM,* and *SCI* for thalidomide search, number of references

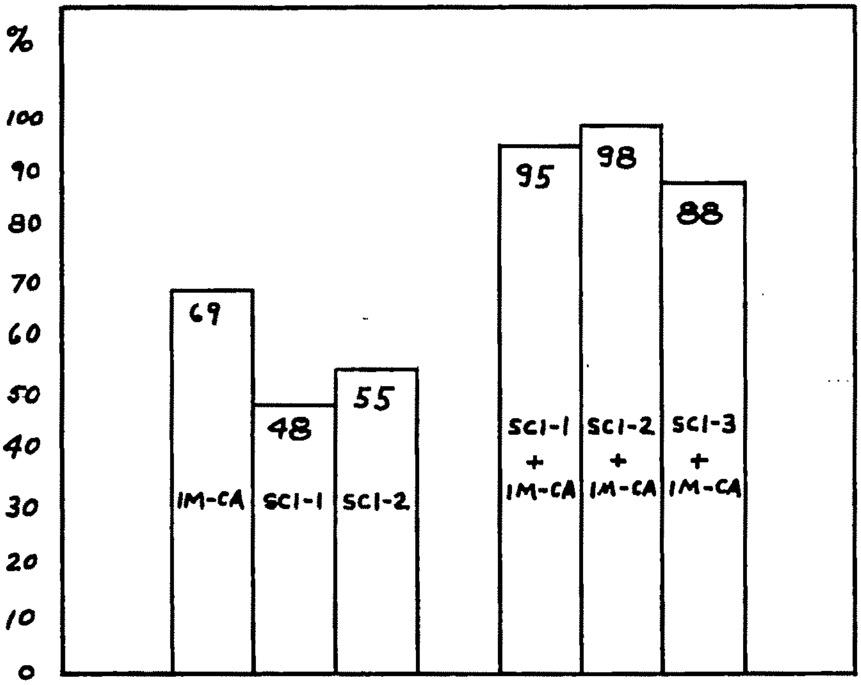


FIG. 5. Relative completeness of *CA-IM, SCI*-1, *SCI*-2, and *SCI*-3 searches. Composite total of 632 references is taken as 100%. Search time for single searches at left was 14.6 hours; for the three combined searches at right, it was 29.2 hours.

TABLE 4. Analysis of search product according to various attributes

								Reason not found in *OA**				Reason not found in *IM**			
	Total	*CA*	*IM*	*OA-IM*	*SOI-1*	*SOI-2*	*SOI-3*	*O*	*J*	*L*	*S*	*O*	*J*	*L*	*S*
Total references	632	110	370	429	302	348	203	15	65	8	434	47	31	39	145
Type															
Case history—original study	525	91	320	363	260	304	173	9	49	7	369	41	14	32	118
Review	26	10	7	14	11	15	7	6		1	9	4	5	7	3
Patent	9	9	0	9									9		
Editorial	24	0	18	18	7	7	8		3		21	1	1		4
Miscellaneous	48	0	25	25	24	22	15		13		35	1	2		20
Field															
Chemical	31	30	4	31	5	3	4			1		5	16	1	5
Clinical	425	8	273	277	194	240	143	1	65	1	350	30	8	16	98
Pharmacological	176	72	93	121	103	105	56	14		6	84	12	7	22	42
Year															
1964	116	24	42	59	68	75	52	10	3	8	71	13	6	39	16
1963	200	44	159	176	53	66	5	2	34		120	6	8		27
1962	207	17	127	136	98	126	66	1	20		169	11	7		62
1961	62	9	14	21	52	50	50		2		51	9	4		35
1960	22	10	14	20	10	10	10	1	2		9	2	4		2
1959	10	2	4	5	9	9	8	1	3		4	5			1
1958	8	1	4	5	7	7	7				7	1	1		2
1957	3	2	2	3	1	1	1				1		1		
1956	4	1	4	4	4	4	4		1		1				
Language															
English	320	41	154	168	211	225	149	12	17	4	246	25	10	26	105
German	146	27	90	106	72	92	48		14	2	103	19	6	10	21
French	38	11	28	33	12	13	6	2	8	1	16	2	1	2	5
Italian	36	9	30	34	2	3		1	5		21	1			5
Other	92	22	68	88	5	15			21	1	48		14	1	9
Origin															
United Kingdom	163	9	62	65	118	128	92	2	5	1	146	4	1	7	89
Germany	121	22	78	90	63	76	39	1	13	1	84	17	6	4	16
United States	108	18	60	64	77	75	45	7	12	3	68	16	7	16	9
Italy	37	10	30	35	2	3		1	5		21	1			6
France	30	9	23	26	8	10	3	1	8		12	1			6
Switzerland	28	6	13	17	15	20	13	3		2	17	3		10	2
Japan	25	10	15	24	1	2	1		10	1	4		9	1	
Other	120	26	89	108	18	34	10		12		82	5	8	1	17
Times cited															
4–24	66	14	44	46	64	65	52	2	3		47	4			18
3	37	5	22	23	34	36	20	1			31	6	1		8
2	70	9	36	37	59	64	30	1	4	1	55	4	1		29
1	159	14	66	75	92	125	58	3	11	2	129	17	9	4	63
0	300	68	202	248	53	58	43	8	47	5	172	16	20	35	27
Most productive reviews															
From:	8	2	1	2	7	8	7	2		1	3	1	2	4	
Lancet	69		25	25	54	61	38				69				44
Brit. Med. J.	62		18	18	48	51	42				62	1		1	42
Arzneimittel-Forsch.	18	11	10	16	12	13	10	1			6	4		3	1
Can. Med. Assoc. J.	14		13	13	4	6	2				14	1			
Deut. Med. Wochsch.	13	1	12	12	7	9	5				12				1
Med. Klin. (Munich)	11	2	7	7	7	8	5				9	4			
Med. Welt	11		8	8	8	8	4				11	2			1
Muench. Med. Wochsch.	10		6	6	5	6	4				10	2			2
Am. J. Obstet. Gynecol.	9		3	3	6	6	2				9	1			5
Nature	8	1	5	5	5	5	3			1	6				3
Science	8	3	4	4	8	8	3	2			3	1		2	1

* *O*, other heading; *J*, journal not covered; *L*, indexed later; *S*, "selective coverage."

10% for *SCI*-2). Pharmacological articles received 41% coverage from *CA*, 53% from *IM*, and 60% from *SCI*-2. Coverage for clinical papers was 2% for *CA*, 64% for *IM* and 57% for *SCI*-2.

IM does not routinely include case histories and studies that appear in the form of letters to the editor. Since the most timely information on drug side effects and toxicity is very often found in short communications rather than in formal papers, an index that covers this material should not be ignored when conducting a drug search.

Most of the anonymous material (editorials) was located through *IM* (75%). One of the most important publications in the drug field, the *Annual Review of Pharmacology*, was not covered by IM until 1965. An index that does cover this publication must necessarily be considered valuable to a drug search.

SCI did not locate any of the nine patents that were located through *CA*. *SCI* did not locate papers in the less common languages as well as did the conventional indexes. While *CA-IM* produced 53% of the English articles and *SCI*-2 70%, and while *CA-IM* located 73% of the German articles compared to 63% for *SCI*-2, for all other languages combined the corresponding figures are 94% for *CA-IM*, and 19% for *SCI*-2.

In some early articles referring to what were later known as toxic effects, the relationship to the drug was not known at the time of reporting, nor was exposure to the drug mentioned, so the article could not possibly have been indexed under the drug name by conventional indexes. However, later reviewers who followed up these cases established that the drug was, or could have been, responsible and incorporated this information into their reviews. For example: Cases of acute myxedema of unknown cause were reported by both Kendall and Hausmann. It was later established that these cases were probably due to thalidomide, which is mentioned in review articles by Gerarde and Mellin. The Gerarde review was located *only* through *SCI*.

The 66 most heavily cited papers found on thalidomide received only 67% coverage from *IM* as compared with 99% coverage from *SCI*-2.

That *SCI*'s coverage is more prompt than *IM*'s can be seen from the number of articles located for 1964 (See Table 4): *IM* located 36% of the 1964 articles compared to 65% for *SCI*-2. *IM* indexed 39/116 of the 1964 articles *after* 1964, almost as many as it indexed *during* 1964 (42/116).

Appendices

Appendix A gives a complete list of language and origin counts. Appendix B is a list of all journals encountered in the search bibliography sample of 632 articles, with the index sources that located them. Appendix C lists all 632 articles on thalidomide with the sources through which they were located. Appendix D shows a sample page of the Citation Index. Appendix E shows a sample page of the Source Index. Appendix F is a tabulation similar to Table 4 characterizing the papers found *only* by *SCI*. Appendix G is a similar tabulation characterizing the papers that were not cited. These appendices are available from the ADI Auxiliary Publication Service.

• Discussion

SCI was expected to appear to rather poor advantage, compared with subject indexes, on a compound name search because there were less than the usual number of terminology problems to reduce effectiveness and efficiency when searching the conventional indexes. For that reason any comparative merits of *SCI* for locating references as specified in the results and conclusions from this experiment are considered conservative.

It must be recognized that this experimental search was only one of many kinds of searches which might have been done for illustration. There is no basis for comparing these results with those of workers cited in the introduction because there is no correspondence in methodology. The results presented here cannot be generalized in any way, but they represent an attempt toward an explicit, unambiguous case study that has brought to light a number of facts about each of the tools studied, and how they compare with one another.

It remains to be seen if the same pattern appears with other compounds, other types of search questions, and other subject areas.

• Conclusions

For this search *SCI* and conventional indexes could be used together profitably; each produced a large number of references not to be found in the other. For this drug search *SCI* was not appreciably less efficient as a retrieval tool than were *CA* and *IM;* for short time intervals, it was more efficient.

More general conclusions must await further investigation.

• Acknowledgments

It is with gratitude that I acknowledge the valuable advice of Mrs. Claire K. Schultz, Senior Research Associate, Institute for Advancement of Medical Communication, throughout the duration of this project.

Thanks are also extended to Dr. Irving H. Sher, Director of Research, Institute for Scientific Information; to Miss Winifred Sewell, Head, Drug Literature Project, National Library of Medicine; and especially to Dr. Richard H. Orr, Director, Institute for Advancement of Medical Communication.

References

1. GARFIELD, E., Science Citation Index, a New Dimension in Indexing, *Science,* 144 (No. 3619):649–654 (1964).
2. MARTYN, J., An Examination of Citation Indexes, *Aslib Proceedings,* 17 (No. 6):184–196 (1965).
3. *Science Citation Index,* Institute for Scientific Information, Philadelphia, 1964.
4. LYKOUDIS, P. S., P. R. LILEY, and Y. S. TOULOUKIAN, Analytical Study of a Method for Literature Search in Abstracting Journals, *Proceedings, International Conference on Scientific Information,* Washington, 1958, 1:351–375 (1959).
5. WALDHART, T. J., A Preliminary Analysis of the Science Citation Index, Thesis, University of Wisconsin (1964).
6. GARFIELD, E., I. H. SHER, and R. J. TORPIE, *The Use of Citation Data in Writing the History of Science,* Institute for Scientific Information, Philadelphia, 1964.
7. BAKER, L. V., A Technique to Discern Patterns of Alkaloid Research in Soviet Block Countries, Thesis, Drexel Institute of Technology, 1966.
8. *Effective Use of the Science Citation Index: A Programmed Text,* Institute for Scientific Information, Philadelphia, 1964.
9. OHLMAN, H., Low Cost Production of Marginal Punched Cards on Accounting Machines, *American Documentation,* 8:123–126 (1957).

A Rationale for Attacking Information Problems

The "systems" approach to information system problems is suggested, wherein problems arising from information origination, processing, and utilization—and alternative solutions to the problems—can be viewed as an entirety rather than piecemeal. Information utilization problems involve sociopolitical considerations (e.g., "wants" vs. "needs" of users), economic values of information, and the more objective considerations of timeliness, quality, and format requirements placed upon information services or products. Quality is encompassed by the factors of specificity, completeness, and relevance. Information processing is shown to consist of seven distinct "unit processes," which may be combined in only nine different ways, thus defining nine possible types of information systems. The "unit processes" employed interact strongly with each other and with user requirements. Information origination—specifically the increasing ratio of "dross" to "ore"—is stated to be the single major information problem for which rational means of attack are not apparent at present.

EUGENE WALL†

Lex-Inc.
Rockville, Maryland

• Introduction

This paper attempts to develop a rationale against which information problems can be viewed and within which the problems can be defined and the solutions to the problems explicitly delimited as to generality.

All problem-solving endeavors can be attempted at a number of different levels of generality, and the problem solutions developed by such endeavors are (usually) generally applicable only to the extent that the problem-solving effort was itself generalized. That is, a problem can be narrowly defined by excluding the apparently less-central variables that might affect the validity of the problem solution. Then, unless the excluded variables are in fact constants, the problem solution may well be invalid in any general sense. The situation is even more serious if the excluded variables are *unknown.* Under these circumstances, even the *limits* of problem solution validity cannot be recognized.

In engineering work, the approach to problem solution via excluding and perhaps *not even defining* the less-central variables has been called by some "methods engineering." This approach solves specific problems without regard to their interactions. Thus each problem solution may be optimal for a part of the "problem complex," but the *combination* of solutions is usually much less-than-optimal for the "complex" as a whole. That is, suboptimization has been achieved.

A different approach has come to be called "systems engineering." This approach (*1*) requires the identification of as many variables as possible (irrespective of their apparent degree of centrality to the problem at hand), the evaluation of their centrality, and the explicit choosing of the more central variables for consideration in problem-solving. Usually more variables are chosen in the systems approach than in the methods approach; i.e., both a broader and more detailed view is taken of the problem. Problem solutions developed via the systems approach are thus more generally applicable than those developed via the methods approach; further, the limits of applicability of the solutions are clearly defined.

It is postulated that to date the methods approach has largely been followed in solving information problems. For example, we have carried out indexing effectiveness studies without a knowledge of what effectiveness means in terms of users' needs. Then we have studied users' needs without regard to the separation of needs from "wants"—"wants" based upon habit, ignorance of what could be had, etc. Many similar situations could be cited. It would seem that the time has come to apply the sys-

† This paper was prepared while the author was an employee of Information Dynamics Corporation, Reading, Massachusetts.

tems approach to the solution of our science information problems.

The first step of the systems approach, as noted, is to identify as many of the pertinent variables as possible, thereby permitting consideration of their degree of centrality to any problem at hand. In information work, there are many problems and many variables. This paper attempts to set forth the variables involved, as a first step toward defining a *network* of variables, different *parts* of which are applicable in varying degrees to different information problems. In fact, the problems facing us are definable in terms of interacting variables; i.e., the interaction of variables creates problems.

As a second step toward defining this network of variables, we attempt herein to organize the variables in a rational manner, thus to make easier the detection of interactions among them and, as a consequence, the definition of specific problems on a rational basis. Except for a few examples, the detection of interaction of variables and the definition of specific problems is not attempted herein. Suffice it to say here that interactions *may* take place among any of the variables to be described later in this paper. The actual existence of such interactions, and defining their importance, is a task for further effort.

The principal problem areas, with their respective variables, are connected intimately with the *communication process.* All *processes* have an *input,* a *processing* operation, and an *output.* It is with these three phases of the communication process that the three problem areas are associated: input, processing and output. We might call them information "origination," information "processing," and information "utilization"—the three areas of issue in this field. Because the *end* to be achieved is that of purposeful information *utilization,* we shall examine the issues in reverse order: utilization, processing, and (finally) origination. For each of these issues, we shall attempt to set forth a framework for problem definition purposes.

• Information Utilization

General

If the objective of information processing activities is the *utilization* of information to improve the cultural or material lot of mankind (or of a segment thereof), then it is essential that the *true needs* of information users be ascertained. By "true needs," we mean those needs which would exist *if* economic and sociopolitical factors were not operational. It must be recognized that such factors, however, *are* operational. Just as mathematicians recognize that it is impossible to trisect an angle by formal techniques, so also must we realize that some apparent needs of users are unrealistic because of their dependence upon economic and sociopolitical influences; and this comment applies to a spectrum of users ranging from those who really want their problem *solved* (not just the information *useful* in solving it) to those who think they need no information at all. Nevertheless, we must determine as best we can what the users' needs *would* be if economic and sociopolitical complications were absent; for only in this way can we have a benchmark against which to measure our progress toward satisfying users' needs—a goal for which we can strive. Accordingly, a discussion of sociopolitical and economic considerations will precede the discussion of objective ("technical") considerations which define the *true* needs of users.

Sociopolitical Considerations

There are undoubtedly cultural, sociological, political, psychological, and other similar, often nonobjective, obstacles to the satisfaction of the *true* needs of users. In order to detect some of these obstacles (as well as more objective problems in the economic and technical areas), scientific disciplines should be encouraged to undertake critical self-examinations. A large self-review (*2*) of the information exchange "culture" in the field of psychology, now four years in progress, is providing some startling insights into the patterns of communication in psychology. If these findings are matched in other fields, thus providing a basis for general conclusions, the sociopolitical complications of scientific communication may be sufficiently defined to permit their rational consideration as part of the communication problem.

It may well be that such nonobjective considerations may make impracticable the development, in the near future, even of economically justifiable systems capable of meeting the *true* needs of users. Instead, it may be necessary to press gradually, in an evolutionary manner, toward the more ideal problem solutions, changing the sociopolitical climate little by little over a long period of time.

Economic Considerations

Today we know essentially nothing about how to measure the *value* of information to users. Under such circumstances, it is impossible to determine what costs can be justified in supplying information to users. The problem is intensified by the probabilistic nature of the value variable.

Several approaches have been probed, tentatively, in attempts to measure the value to the user of information. For example, a number of organizations have shown that their organized information activities, as presently constituted, cost *no more* than if the user searched for information independently, and to the extent desired *by* the user. Such an evaluation, however, gives no consideration to the cost/savings ratio under more rational definitions of user needs or with more optimally designed information services for the user. Another approach is that of the *reductio ad absurdum*—whereby it can be shown that if a scientific investigation costs $20,000, if the report thereon costs $2,000, and if the average report is referred to 10 times over its useful lifetime, then each reference costs about $200 or 1% of the cost of the investigation; surely the average user benefits more than 1% from his use of the report. Such an evaluation is essentially circular, however, and provides only a subjective rationalization of the value of information services.

What is really needed is a way of determining what the situation *would be,* from an economic point of view, if information is not made available; e.g., how profitable a manufacturing plant would be if an information service of specified caliber is available during its design, compared to the reverse situation. This is a very difficult problem, but similar problems, when well defined, have been capable of reasonable solution via operations research techniques, even when probabilistic conditions prevail, as in this instance.

Technical Considerations

General

With respect to true users' needs, three types of technical considerations are pertinent: the performance of information services as concerns their *timeliness, quality,* and *form or format* of information or data supplied to the user.

Timeliness

Certain users need information more quickly than others; therefore, a measure of timeliness requirements (like that of permissible cost) will be of a probabilistic nature. This factor applies either to a current awareness service (i.e., how *current* is the information supplied?) or to a retrospective service (i.e., how *soon* does a search result in an "answer"?).

The principal problem here is that of being able to characterize users (and information-need situations) so that timeliness requirements can be objectively measured or predicted.

Quality of Information Suppplied to Users

General. It is probable that the quality of information required by (and supplied to) users should be capable of definition via three parameters: *specificity, completeness* and *relevance.* These parameters are discussed in more detail in the following paragraphs.

Specificity. Different users require information with different degrees of specificity. Some users may usually require quantitative data (e.g., the boiling point of water under a pressure of 300 psig). Other users may usually require conceptual—even subjective or speculative—information (e.g., a discussion of the likelihood of intelligent life existing elsewhere in our galaxy). Between the limits set by purely quantitative and purely qualitative information, there exists a continuum. Against this continuum each user will exhibit a distribution of interest points, but each user will also probably find that most of the time his requirements for specificity fall within a relatively narrow range.

Completeness. Different users require different degrees of completeness[1] of the information supplied to them (i.e., different degrees of exhaustiveness of retrieval). Hence this factor also exhibits a probabilistic nature. It is not always best to supply the user with *all* documents that answer his need. For example, the user who needs to know the boiling point of water at 300 psig would *prefer* to have a dimensioned number—a temperature—supplied to him. Lacking that, he would like to have one authoritative document in which the desired datum is recorded. He would be quite unhappy to receive a hundred, or even a dozen, documents *even if all* contain the desired datum. On the other hand, the scientist who needs less specific information is also likely to require more completeness of retrieval. Thus *specificity* requirements and *completeness* requirements interact.

Relevance. Different users require different degrees of relevance of the information supplied to them. Thus a probabilistic distribution of requirements exists here, also. In general, the more specific the information desired, the greater the need for relevance—the need *not* to be burdened with nonpertinent information and documents, or both. The reverse situation also holds true; the more general the information desired, the less the need for relevance. Similarly, the greater the need for completeness, the less the need for relevance, in that peripherally pertinent information tends to be acceptable and even useful in such instances.

Form or Format of Information or Data

This factor is principally concerned with *ease of use* and is probably the easiest-to-measure user need. For example, should *data* be displayed as tables, graphs, alignment charts, or equations? Should the "raw," non-reduced data be included? How should a display make plain the constant (yet specific) conditions under which the data were collected—conditions which, if changed, might change the data values observed? Should the data be described (e.g., possibly for announcement purposes) by an abstract? Should the data and/or abstracts be indexed, and if so, how deeply? How should different set of data be grouped or categorized so that proximity of related data sets (from the users' point of view) is maximized?

A similar set of considerations apply with respect to textual information. What level or levels of surrogation should be made available to the user—full text, sum-

[1] Sometimes called "recall."

mary, informative abstract, indicative abstract, notation-of-content, citation, title, or document number? What "depth" of indexing is required, if any? How should documents or their various surrogates be grouped or categorized?

With respect to indexes alone, there are important considerations (other than that of indexing "depth") related to format, type of index entry, and character of index terminology. What format is optimal? Should a particular index be a classification or an alphabetical index (modified or not by subheads)? Should index entries be merely document numbers; or should they include citations, notations-of-content, or even abstracts? Should the index terms be complex (e.g., classification notations) or simple (e.g., uniterms) or something in between (e.g., subject headings of varying degrees of complexity)?

With respect to format generally, should the data, information, surrogates, indexes, etc., be printed; and if so, on pages or on cards, and in what arrangement on page or card? Should these materials be full size or microform (continuous or discrete—e.g., microfiche, positive or negative image, transparent or opaque) or something between full size and microform? Or should these materials be recorded on other types of media, such as magnetic tape, drums, matrices, discs, chips, etc.? If the recording medium is something other than full-size printed text, what sorts of display equipment with what speeds of access and operation are required?

• Information Processing after Origination but Prior to End-Use

General

It is in this area that nearly all investigatory work has been concentrated for centuries—unfortunately, too often with too little consideration for input (see following paragraphs) or output (see preceding paragraphs) restraints. It is thus no small wonder that an information crisis exists today; it is surprising that even a worse crisis has not developed. This is not to denigrate the effort that has been expended with respect to information processing. Such effort is *necessary;* it is, however, *not sufficient*—and considerations of input and output should exert a much greater influence on information processing activities. This is particularly true with respect to the needs of users (and the statistical distribution of those needs), as already described.

Despite this caveat, let us examine the overall field of information processing. A number of questions (but not all appropriate questions, undoubtedly) spring to mind immediately. What are the "unit processes" of this "middleman" operation? How may these "unit processes" be assembled into various systems in order to serve the needs of various users, taking into account also the characteristics of input information? How may the details of the "unit processes" be varied, and for what reasons, considering also the interactions among the processes both within and among systems? How may systems each designed to serve a certain type of need by processing a certain type of input, best be interconnected? To what degree, and why, should such systems be centralized or decentralized?

Within the processing area, we can detect seven distinct "unit processes," each of which may be widely varied in detail and in scope. These are the processes that we shall tag with the *general* terms of acquisition, surrogation, announcement, index operation, document management, correlation, and vocabulary control. From one or more of these seven processes can be constructed every information or data system, excluding those consisting of *direct communication* between the originator and user of information. These unit processes are briefly discussed in the next section.

The "Unit Processes" of Indirect Communication

Acquisition

This process also encompasses evaluation and/or selection of documents or data for input into current awareness, announcement and/or retrieval systems, into correlation or data reduction processes, or for use without further processing; it also includes descriptive cataloging and duplicate checking.

Surrogation

This process includes abstracting, indexing, data reduction, or the like. It should be noted that information *not* present in the full-text material, or raw data, cannot be reduced to surrogates. This surrogation is no substitute for serendipity, for reasoning by analogy, or for creative thought generally.

Announcement

This process encompasses the assembly, production and distribution generally or specifically (i.e., selective dissemination of information) of announcement media for documents, data tables, etc.

Index Operation

This process includes input of index information into a physical medium and the searching of that medium routinely or on demand (including the "negotiation" and formal formulation of inquiries), and the provision of outputs consisting of actual data or of references and/or other useful document surrogates (e.g., abstracts).

Document Management

This process encompasses the physical storage and retrieval (based upon "addresses" only) of documents or data and the reproduction and inventory control of such items.

Correlation

This process creates generalized information from multiple, other, more specific items. The creation of state-of-art reviews is an example of the correlation process. Analysis to determine the significance of reduced data is another example. In short, correlation is a creative activity and differs from the research process (as it is usually defined) only in that correlation per se requires no material experimental activity on the part of the correlator.

Vocabulary Control

This process includes all operations useful in bringing into conjunction the vocabularies of originators, processors, and users of data and information, to the end that the communication channel will be as effective and yet noise-free as possible. The physical impedimenta with which the vocabularly control process is most often concerned are the various types of authority files (including but not limited to *subject* authority files), together with their syndetic structures and conventions for updating and use.

Systematization of the "Unit Processes"

If vocabulary control is considered as a general function of all systems, to be applied (or not) to the required degree, then a limited number of *basic* information systems are possible. Figure 1 indicates that only eight basic systems are possible, excluding that system (numbered "zero") consisting of direct communication between the originator and user of information. That is, only eight unique paths through the "unit processes" from origination to end-use are possible.

External to a *given* system, of course, things can get much more complicated. Interfaces *between* information systems can occur, bringing about (in effect) loops in the pattern. Two principal types of *functional* interfaces can exist. The first, indicated by a dotted line, may be termed *reorigination*. The second, indicated by dashed lines, may be termed *reacquisition*. In addition, interfaces with respect to information or data coverage and clientele servicing may exist.

Designing the Information System

Ideally, when users' needs have been defined, the required combination of "unit processes" should be assembled to create a system, and the necessary input information or data should be located and processed for use. Because the details of the "unit processes" interact with each other, each such process must be designed as a *part of the system*—not independently. For example, the surrogation process of indexing for announcement purposes only should probably differ markedly from the process of indexing for retrospective retrieval purposes. Again, the work done in developing variants of the "unit processes" and in meshing them into systems has been valuable and *necessary*. It must be reiterated, though, that such work is *not sufficient* to solve all the problems that face us. Work on improving the processing of information, between its origination and use, must take place with full cognizance of the restraints applied by origination and use.

In conclusion with respect to information processing, we note that not only the environmental variables (e.g., users' needs) but also the available operational techniques all interact strongly.

• **Information Origination**

Of the three areas of concern, that of information origination has until very recently consumed more of our effort, and has received *less* of our rational attention, than has either of the other two areas. For example, we have spent much time and money in *publishing* information, but relatively little in deciding what *should* be published. It is apparent that much (perhaps most) of the information being originated should never go "on the record," that a great outflowing of trivia or duplicative information is being published or disseminated in one manner or another. Such a proportion (perhaps preponderance) of dross in our raw material makes recovery of valuable materials much more difficult and costly. The problem facing us is: "How can we effect birth control of information such that we exclude from the communication pattern as much of the trivia as possible *without* excluding useful information?" Several approaches are apparent, but none of them seem to have the full potential which we desire.

We can, of course, continue to work to make available (to each potential originator of information) the pertinent scientific or technical information developed in the past, in the hope that the potential originator would be so knowledgeable that he would not undertake the pseudocreative effort to originate trivia or duplicative information. This may be a forlorn hope, in view of the all-too-frequent practice of re-reporting one's *own* work. All other things being equal, human beings may still attempt to build their own professional stature by whatever means they can.

Another approach would be to insure that all work is well evaluated (before its publication or dissemination) by knowledgeable yet disinterested referees. Here we face the problem of finding ways of inserting referees into each of the many (often devious) routes that a persistent originator might employ to inject his trivia into the communication pattern. We also face the problem of having truly knowledgeable referees, although

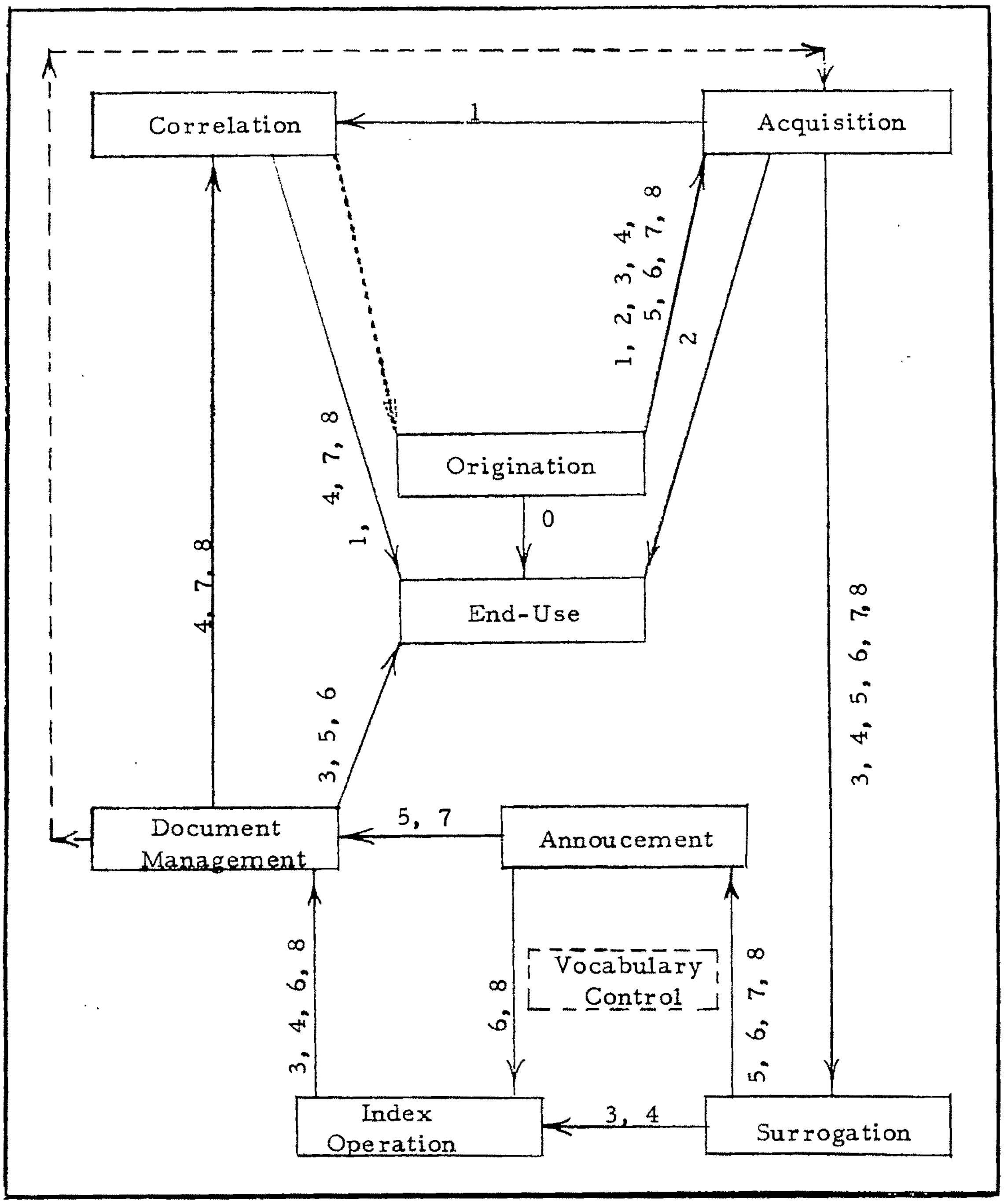


FIG. 1. Basic information systems and system functions

this obstacle may someday be overcome by giving to each referee not only the candidate informational item but also all pertinent information on the same subject which has been developed in the past. Finally, under such circumstances, will not the referees need compensation, and if so, who will pay?

The most effective solution to the information birth control problem—and the most difficult solution to implement—would be to attack the problem at the source, at the originators themselves. We should like to find a way of making the originators actively *want t* *avoid* reporting trivia. At this time, however, I see n positive incentive for originators to be so altruisti There seems to exist only the negative incentive o ridicule, and ridicule is relatively ineffective becaus there will always be uninformed readers who will gran to the originator of trivia at least some reputation o acclaim.

It seems likely that the information origination problem will be with us for some while.

Summary

Many attempts made to date to solve information problems have been either ineffective or else not demonstrably optimally successful because insufficient attention has been given, during the problem-definition stage, to a sufficient number of the important variables involved. In this usual "methods" approach, the interactions between only two (or among a few) variables are investigated, without control or perhaps even without observation of other important variables, and problem solutions are developed only in terms of the controlled variables. The variables are usually probabilistic in nature (i.e., statistical distributions exist) and all may interact strongly. It is therefore suggested that the "systems" approach should be employed, wherein all important interacting variables are considered and measured and their measurements correlated during problem definition and solution.

The important variables are described in three categories: information *utilization*, *processing*, and *origination*. In the *utilization* category, sociopolitical (and, to some extent, economic) considerations constitute variables that are still difficult to define and to measure. Yet the distinction must be made, before problem definition can be effective, between true user *needs* and apparent user "*wants*," the latter the result of habit, complacence, ignorance, fear, or lack of motivation. Similarly, the *value* of information to users must be quantified. Beyond sociopolitical and economic variables, there are technical variables of user needs with respect to time (or timeliness) and quality of service provided. The latter requirement variable can be subdivided into the factors of relevance, completeness, specificity, form, and format.

The *processing* variables are related to functions that may be performed by an information facility—i.e., acquisition (and associated activities), surrogation (e.g., abstracting and indexing), announcement, index operation (including input and searching), document management (including storage, dissemination, and related processes), correlation (the creation of generalized information from multiple, more specific inputs), and vocabulary control. It is shown that the possible combinations of these processes permit only eight basic types of information facilities to exist.

Information *origination* appears to be the least well-defined problem area, and further definition of its variables is required. The problem may be broadly stated as that of excluding trivia and other information of low utility from the communication channels while still accepting valuable information. This statement, however, is much too general to serve operational purposes in information research endeavors.

References

1. Goode, H. H., and R. E. Machol, *System Engineering*, McGraw-Hill, New York, 1957.
2. Garvey, W. D., and B. C. Griffith, *An Overview of the Structure, Objectives, and Findings of the American Psychological Association's Project on Scientific Information Exchange in Psychology*, American Psychological Association, Washington, D. C., August 1963.

Distribution of Indexing Terms for Maximum Efficiency of Information Transmission

A function was developed for the optimum distribution of indexing terms by the number of postings. This makes it possible to transmit information with maximum efficiency. The comparison of the actual distribution of the term groups with the calculated optimum distribution provides an objective measure for evaluating any indexing system with respect to its efficiency as information transmission channel.

PRANAS ZUNDE and VLADIMIR SLAMECKA

School of Information Science
Georgia Institute of Technology
Atlanta, Georgia

Organization of indexing data in a manner that permits retrieval of the greatest amount of information is a premise fundamental to an efficient performance of all information storage and retrieval systems. In this context an index may be considered a channel linking the information store and the user or searcher. The problem may be viewed as that of organizing information store—that is, the indexing "terms" (subject descriptors) and their "postings" (document identification numbers, for instance)—so as to make a maximum use of channel capacity, permitting the transmission of information from the store to the user with maximum efficiency.

For a given system, indexing terms can be grouped by the number of postings they carry, to form a frequency spectrum characteristic of that system at a particular time. This paper reports the development of a distribution function of term groups by number of postings which allows information transmission with maximum efficiency, and it proposes a measure of evaluation of a system's efficiency in this respect.

The solutions are developed subject to three assumptions. First, no distinction is made between useful and useless information; only the amount of information, not its subjective value, is considered. Second, the possible effects of indexing language and of the function and form of index terms are not taken into consideration, as they are likely to vary from one vocabulary to another. Third, the channel is assumed to be a noiseless one; that is, if term T_1 is addressed, the probability of retrieving it with all its postings is equal to 1.

The efficiency of a statistical information transmission model is defined as

$$\eta = \frac{I(T,Y)}{C} \tag{1}$$

where $I(T,Y)$ is the transinformation and C is the channel capacity. The channel capacity is in turn defined as

$$C = [I(T,Y)]_{\max} = [H(T) - H(T/Y)]_{\max} \tag{2}$$

In this equation, $H(T)$ is the source entropy, which in our case is the entropy derived from the relative frequencies of indexing terms by the number of postings, and $H(T/Y)$ is the conditional entropy or average uncertainty given that a particular symbol has been received as to the symbol which was transmitted.

Since the channel was assumed noiseless, the conditional entropy

$$H(T/Y) = 0$$

and the transinformation

$$I(T,Y) = H(T)$$

Thus

$$C = [H(T)]_{\max} \tag{3}$$

In information theory, the entropy or measure of uncertainty of a complete finite scheme is given by

$$H(T) = -\sum_{t=1}^{n} p(t) \log p(t) \tag{4}$$

where $p(t)$ is the probability of the occurrence of the event T (in our case, the probability of term group T_t having t postings).

From Eq. (1) we see that the efficiency of an index system is highest when transinformation is equal to the channel capacity. Since in our case the channel capacity is equal to maximum source entropy, the problem is to find such a frequency distribution of term groups by number of postings which produces maximum source entropy. The solution is subject to two constraints:

$$\sum_{t=1}^{n} p(t) = 1 \tag{5}$$

and

$$\sum_{t=1}^{n} tp(t) = \bar{t} = \text{const} \tag{6}$$

The first constraint states that the sum of the relative frequencies is equal to one (viz., we have a complete finite scheme). The second constraint states that there is a fixed amount of postings in a given system, expressed as the average number of postings per term.

Using the method of calculus of variation and Lagrange undetermined multipliers, we can write:

$$\delta H(T) = -\Sigma[\ln p(t) + 1]\delta p(t) = 0 \tag{7}$$

$$\alpha\Sigma\delta p(t) = 0 \tag{8}$$

$$\beta\Sigma t\delta p(t) = 0 \tag{9}$$

Adding Equations (7), (8), and (9), we get

$$\Sigma[\ln p(t) + \alpha + \beta t]\delta p(t) = 0 \tag{10}$$

hence

$$\ln p(t) + \alpha + \beta t = 0 \tag{11}$$

or

$$p(t) = e^{-\alpha}e^{-\beta t} \tag{12}$$

To find the Lagrange constants $\propto$ and β, we turn to the constraints equations. Substituting Eq. (12) into Eq. (5), we obtain

$$\sum_{t=1}^{n} e^{-\alpha}e^{-\beta t} = 1 \tag{13}$$

hence

$$e^{-\alpha} = \frac{1}{\sum_{t=1}^{n} e^{-\beta t}} \tag{14}$$

Substituting the expression for $e^{-\alpha}$ into Equation (12), we obtain

$$p(t) = \frac{e^{-\beta t}}{\sum_{t=1}^{n} e^{-\beta t}} \tag{15}$$

Substituting the expression in Eq. (15) for $p(t)$ into the second constraint Equation (6), we find

$$\frac{\sum_{t=1}^{n} te^{-\beta t}}{\sum_{t=1}^{n} e^{-\beta t}} = \bar{t} \tag{16}$$

The summation limits are the lowest and the highest number of postings under any one term in our vocabulary (i.e., 1 and n). This range of postings is, obviously, bounded; we can, however, extend the upper summation limit to infinity by simply considering the remainder in our series as the sum of terms corresponding to the frequency of terms with "$k+1$ and more" postings, provided k is large enough.[1] Both series in the numerator and denominator are convergent series for $\beta > 0$. Under this assumption, the sum of the series in the denominator is easily established. Thus

$$\sum_{t=1}^{\infty} e^{-\beta t} = \frac{e^{-\beta}}{1 - e^{-\beta}} \tag{17}$$

The limit of the series in the numerator can be found as follows:

$$\sum_{t=1}^{\infty} te^{-\beta t} = \frac{\partial}{\partial\beta}\left[-\sum_{t=1}^{\infty} e^{-\beta t}\right] = \frac{\partial}{\partial\beta}\left[-\frac{e^{-\beta}}{1 - e^{-\beta}}\right] = \frac{e^{-\beta}}{(1 - e^{-\beta})^2} \tag{18}$$

Hence

$$\bar{t} = \frac{\dfrac{e^{-\beta}}{(1 - e^{-\beta})^2}}{\dfrac{e^{-\beta}}{1 - e^{-\beta}}} = \frac{1}{1 - e^{-\beta}} \tag{19}$$

and

$$\beta = -\ln\left(1 - \frac{1}{\bar{t}}\right) \tag{20}$$

By substituting the values of β into the equation, we obtain the frequency distribution function that maximizes the source entropy:

$$p(t) = \frac{\left(1 - \frac{1}{\bar{t}}\right)^t}{\sum_{t=1}^{\infty}\left(1 - \frac{1}{\bar{t}}\right)^t} \tag{21}$$

The limit of the convergent series in the denominator is easily obtained as

$$\sum_{t=1}^{\infty}\left(1 - \frac{1}{\bar{t}}\right)^t = \bar{t} - 1 \tag{22}$$

and finally we write

$$p(t) = \frac{\left(1 - \frac{1}{\bar{t}}\right)^t}{\bar{t} - 1} \tag{23}$$

Example. For an information storage and retrieval system with 1,000 documents, each of which is indexed by 22 terms on the average, with a vocabulary of 2,000 terms and the average number of posting per term $\bar{t} = 11$, the optimum frequency distribution of term groups to produce the maximum average amount of information per term would have approximately 182 terms with one posting, 165 terms with two postings, 146 terms with three postings, and so on. The percentage

[1] The effect of the approximation error should be evaluated in each particular instance. For small $\bar{t}$ and β ($\beta \ll 1$), it might be significant.

of terms which should have 100 or more postings can be calculated as follows:

$$\sum_{t=100}^{\infty} p(t) = \sum_{t=100}^{\infty} p(t) - \sum_{t=1}^{\infty} p(t)$$

$$= \sum_{t=1}^{\infty} \frac{\left(1-\frac{1}{\bar{t}}\right)^{t}}{\bar{t}-1} - \sum_{t=1}^{\infty} \frac{\left(1-\frac{1}{\bar{t}}\right)^{t}}{\bar{t}-1}$$

$$= 1 - \frac{1}{\bar{t}-1}\left[\frac{1-\frac{1}{\bar{t}}}{\frac{1}{\bar{t}}} - \frac{\left(1-\frac{1}{\bar{t}}\right)\left(1-\frac{1}{\bar{t}}\right)^{100}}{\frac{1}{\bar{t}}}\right]$$

$$= (\bar{t}-1)\left(1-\frac{1}{\bar{t}}\right)^{100} = 0.00072$$

Thus, in this example of a system, 0.072% of terms (between one and two terms) should have 100 or more postings. The percentage of terms with high numbers of postings clearly depends on the average number of postings per term: the greater this average, the higher the percentage of heavily posted terms.

An immediate use of the concept of term group distribution is in the evaluation of existing information systems. Figure 1 compares the actual and optimal term group distribution curves of two information systems, that of the Defense Documentation Center (in 1960) and a private experimental one (*1*). Although the actual distribution curves of both systems (solid lines in the graph) differ from their optima (broken lines)—implying a less-than-efficient use of channel capacity—the

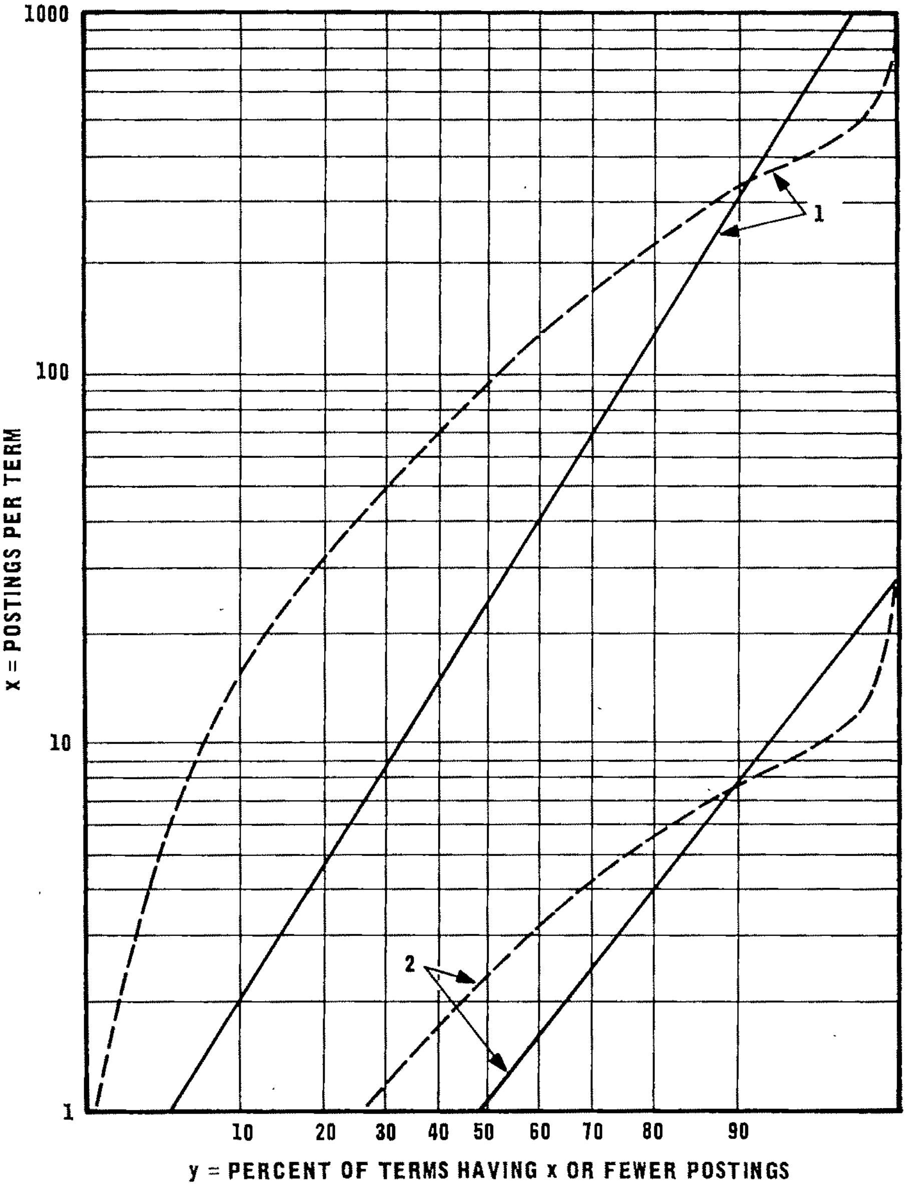


FIG. 1. Actual vs. optimum term-group cumulative distribution in two information systems; 1 = Defense Documentation Center System (1960); 2 = private experimental system (1960)

comparison of the efficiency coefficient will show that the deviation is considerably more serious in the experimental index.

To calculate the efficiency coefficient of a given system, we calculate the transinformation $I(T,Y)$ from the given frequency distribution of term groups. Next, the channel capacity for the system can be derived as follows.

It has been shown (*2*) that, for a discrete random variable with

$$\Sigma p(t_i) = 1 \qquad \text{and the mean} \qquad \Sigma t_i p(t_i) = \bar{t}$$

$$H(I) = -\sum_{i=1}^{n} p(t_i) \ln p(t_i) \leqq \ln M(\beta) + \beta\bar{t} \quad (24)$$

with equality if and only if $p(t_i)$ is the maximizing function of the entropy equation. Then

$$M(\beta) = \sum_{i=1}^{n} e^{-\beta t_i} \quad (25)$$

where β is given by

$$\beta = -\ln\left(1 - \frac{1}{\bar{t}}\right) \quad (26)$$

Furthermore, we have already shown that

$$\sum_{i=1}^{n\to\infty} e^{-\beta t_i} = \frac{e^{-\beta}}{1-e^{-\beta}} \quad (27)$$

Substituting the expression for β into the Equation (27), we get

$$M(\beta) = \sum_{i=1}^{\infty} e^{-\beta t_i} = \frac{\exp\left[+\ln\left(1-\frac{1}{\bar{t}}\right)\right]}{1-\exp\left[\ln\left(1-\frac{1}{\bar{t}}\right)\right]} = \bar{t} - 1 \quad (28)$$

Thus, we can rewrite Equation (24) as follows:

$$[H(T)]_{\max} = \ln M(\beta) + \beta\bar{t} = \ln(\bar{t}-1) - \bar{t}\ln\left(1-\frac{1}{\bar{t}}\right)$$
$$= \bar{t}\ln\bar{t} - (\bar{t}-1)\ln(\bar{t}-1) \quad (29)$$

which is the expression for channel capacity.

According to the last equation, the channel capacities of the experimental system with $\bar{t}=3.9$ and of the DDC system with $\bar{t}=148$ would be 2.22 nits and 6 nits, respectively. By interpolating the graph reproduced in Houston and Wall's article (*1*) (since complete data were not available to us), we obtained approximately 1.66 nits actual source entropy (=transinformation in the given core) for the experimental system and 5.4 nits for the DDC system. Hence, the efficiency coefficients are approximately 0.75 (75%) and 0.9 (90%), respectively.

In their article Houston and Wall, having investigated a number of indexes for frequency distribution of postings under the terms, came to the conclusion that the distribution is lognormal for all indexes. Furthermore, they derived an empirical probability distribution function, which is supposed to give the frequency spectrum of all actual indexes, if the sample they investigated was representative enough. The proposed function is:

$$p(x) = \frac{0.17}{x}\exp\left[-\frac{(A\log x - B)^2}{2}\right] \quad (30)$$

where

x = number of terms with x postings (x=1,2,3, . . .)
$A = 2.0 - 0.14 \log_{10} P$
$B = 0.67 \log_{10} P - 2.4$
P = total number of postings

It can be shown that the average source entropy for this type of distribution can be approximated by

$$H(X) = -\sum_{i=1}^{n} p(x_i) \ln p(x_i)$$

$$\cong \left[-\frac{0.17 \ln 0.17\sqrt{\pi}}{\sqrt{2}\,A\log e} + \frac{0.17\sqrt{\pi}\ln 10(0.67\log P - 2.4)}{\sqrt{2}\log e\,(2.0-0.14\log P)^2}\right.$$

$$\left.+\frac{0.17\sqrt{\pi}}{2\log e\,(2.0-0.14\log P)}\right]\left\{\operatorname{erf}\frac{(0.67\log P - 2.4)}{\sqrt{2}}\right.$$

$$+\operatorname{erf}\left[\frac{1}{\sqrt{2}}(2.0\log x_n - 0.14\log P\log x_n\right.$$

$$\left.\left.-0.67\log P + 2.4)\right]\right\} - \frac{0.17\ln 10}{\log e\,(2.0-0.14\log P)^2}$$

$$\exp\{-\tfrac{1}{2}(4.0 - 0.56\log P + 0.0196\log^2 P)\log^2 x_n$$

$$-\tfrac{1}{2}[-2(2.0-0.14\log P)(0.67\log P - 2.4)\log x_n$$

$$+(0.67\log P - 2.4)^2] - \tfrac{1}{2}(0.67\log P - 2.4)^2\}$$

$$+\frac{0.17}{2\log e\,(2.0-0.14\log P)}\{-[(2.0-0.14\log P)\log x_n$$

$$-(0.67\log P - 2.4)]$$

$$\exp\left[-\frac{[2.0-0.14\log P)\log x_n - (0.67\log P - 2.4)]^2}{2}\right]$$

$$-(0.67\log P - 2.4)\exp\left[-\frac{(0.67\log P - 2.4)^2}{2}\right]\} \quad (31)$$

It has been observed that the maximum number of postings per term is approximately $t_{\max}=0.03P$. Substituting this value in Equation (31), we get

$$H(X) = \left[\frac{-0.2793 + 0.5867\log P}{(2.0-0.14\log P)^2}\right]$$

$$\times\left[\operatorname{erf}\left(\frac{0.67\log P - 2.4}{\sqrt{2}}\right)\right.$$

$$\left.+\operatorname{erf}\left(\frac{1.543\log P - 0.14\log^2 P - 0.646}{\sqrt{2}}\right)\right]$$

$$-\left[\frac{0.9014}{(2.0-0.14\log P)^2}\right][\exp(-0.0098\log^4 P$$

$$+0.2162\log^3 P - 1.2852\log^2 P + 1.0268\log P - 0.2076)$$

$$-\exp(0.2245\log^2 P + 1.61\log P - 2.88)]$$

$$+\frac{0.1957}{2.0-0.14\log P}[0.14\log^2 P - 2.8832\log P + 5.446]$$

$$\times\exp(-0.0048\log^4 P + 0.4036\log^3 P - 4.9188\log^2 P$$

$$+15.7019\log P - 14.8294) - (0.67\log P - 2.4)$$

$$\exp(-0.2245\log^2 P + 1.61\log P - 2.88) \quad (32)$$

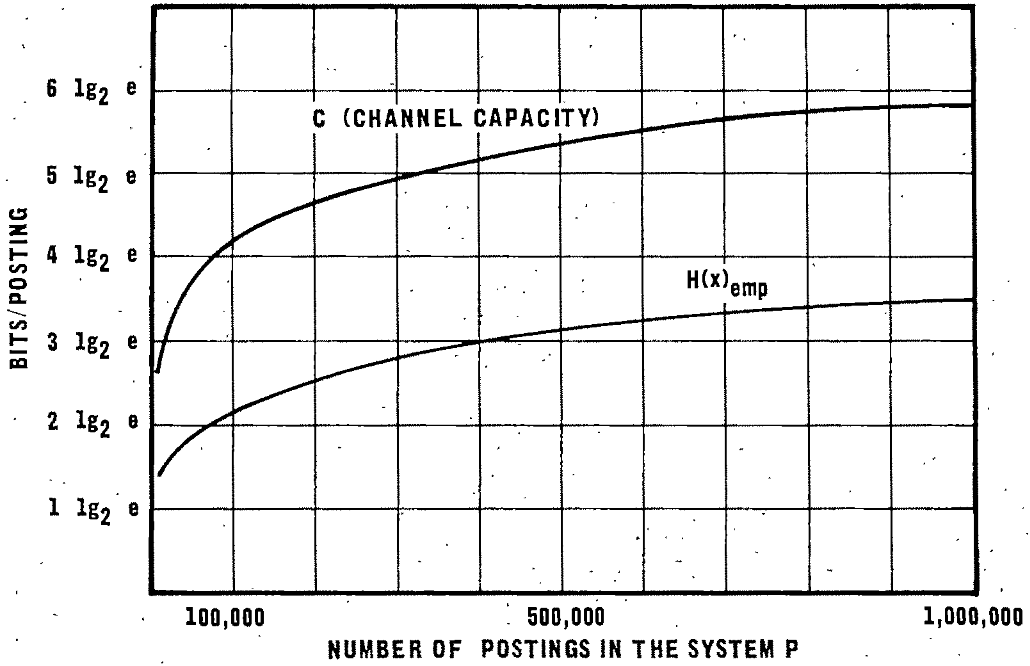


FIG. 2. Actual source entropies $H(X)_{emp}$ and channel capacities C of systems with indexing terms distributed according to Eq. (30)

Houston and Wall also propose in their paper an empirical formula relating the total number of terms in the index with the total number of postings:

$$T = 3{,}300 \log (P + 10{,}000) - 12{,}600$$

where T is the total number of terms and P—total number of postings. The formula is valid for the range of P from 10,000 to 1,000,000. Then

$$\bar{p} = \frac{P}{T} = \frac{P}{3{,}300 \log (P + 10{,}000) - 12{,}600} \qquad (33)$$

Substituting this expression into Equation (29), we obtain the channel capacity as a function of P. On the other hand, from Equation (32) we can calculate the actual source entropy for indexes for various values of P, if the frequency distribution of their postings conforms with Houston and Wall's formula. The corresponding efficiency coefficient will be the ratio of these values.

In Fig. 2, channel capacities and source entropies of indexes that obey the empirical distribution formula of Houston and Wall are plotted as functions of the total number of postings P.

A fuller discussion of the uses of this technique in the design of information systems and in the control and optimization of the indexing and storage apparatus of existing systems is in preparation.

References

1. HOUSTON, N., and E. WALL, The Distribution of Term Usage in Manipulative Indexes, *Am. Doc.*, 15:109 (1964).
2. KULLBACK, S., *Information Theory and Statistics*, Wiley, New York, 1959, p. 34.

Brief Communications

$'s and Secrets

The purpose of this brief commentary is to attempt to provoke the interest of *researchers* concerned with the scientific and technical information problem into considering the impact that $'s and secrets could have on the problem.

"$'s" as used here relate to the proprietary interests of industry in its scientific and technical information, whether resulting from industry funds exclusively, or whether including some public funds. The $'s also relate to the publishing industry, and to copyright.

"Secrets" as used here relate to the responsibility of the Federal establishment to preserve the security of the United States *and* the attendant need to control that scientific and technical information vital thereto.

While these aspects of the scientific and technical information problem (STIP) have, indeed, been recognized, most present STIP research seems to shy away from $'s and secrets. It's easy to understand why, since both present extremely sensitive and delicate problems. In the case of $'s, all sorts of industrial rights are involved. As for secrets, national security itself could be affected. And notwithstanding $'s, secrets, and the STIP, such progress in science and engineering is being made that it may be that the STIP itself has been overemphasized.

If we accept the observation that progress is being made in science and engineering, irrespective of $'s, secrets, and the STIP, then we can hypothesize either that the STIP only *might* make scientific and engineering progress slower or more expensive than optimal, or that the progress that is being made is accomplished by those scientists and engineers who are Information-Haves. That is, technological progress is made both by those people who really have no STIP, and by those who have the necessary access and need-to-know for $'s and secrets. Should this line of reasoning be even remotely correct, then $'s and secrets could be a more important portion of the STIP than those portions that we now emphasize.

Consider the information/data transfer spectrum shown in Fig. 1. The purpose of this figure is to accentuate the impact that release conditions ($'s and secrets) could have on information transfer. Obviously, if the "needer" knows the significant originators, and if he has the necessary need-to-know, he can go directly to the originators and forget the rest of the spectrum (e.g., join the two ends of the figure together). But, possibly many of us don't know all the significant originators, nor do we have all the need-to-know. However, for that information not constrained by $'s and secrets, we soon learn where to approach the transfer spectrum. But what do we do about $'s and secrets?

When the $'s relate to scientific or technological achievements that have industrial market value (e.g., a patent) and if the achievements can be identified, then purchase alone can provide the information. If not, probably there's nothing to be done but to take the chance of duplicating. And in the case of commercial publications, the only solution may be to wait until the information is for sale.

When the scientific and technical information is secret, how do you know it exists, unless of course you are already doing work for one of the Federal agencies? And, even if you are, how do you know that you have access to necessary information produced by some other Federal agency?

Typical questions that arise as a consequence of thinking about $'s, secrets, and the STIP are:

1. Do $'s and secrets have any impact on the progress of science and engineering?
2. As a general rule, is the information that is constrained by $'s and secrets better than the information that is not so constrained?

GUSTAVUS S. SIMPSON, JR., AND JOHN W. MURDOCK
Battelle Memorial Institute
Columbus Laboratories
Columbus, Ohio

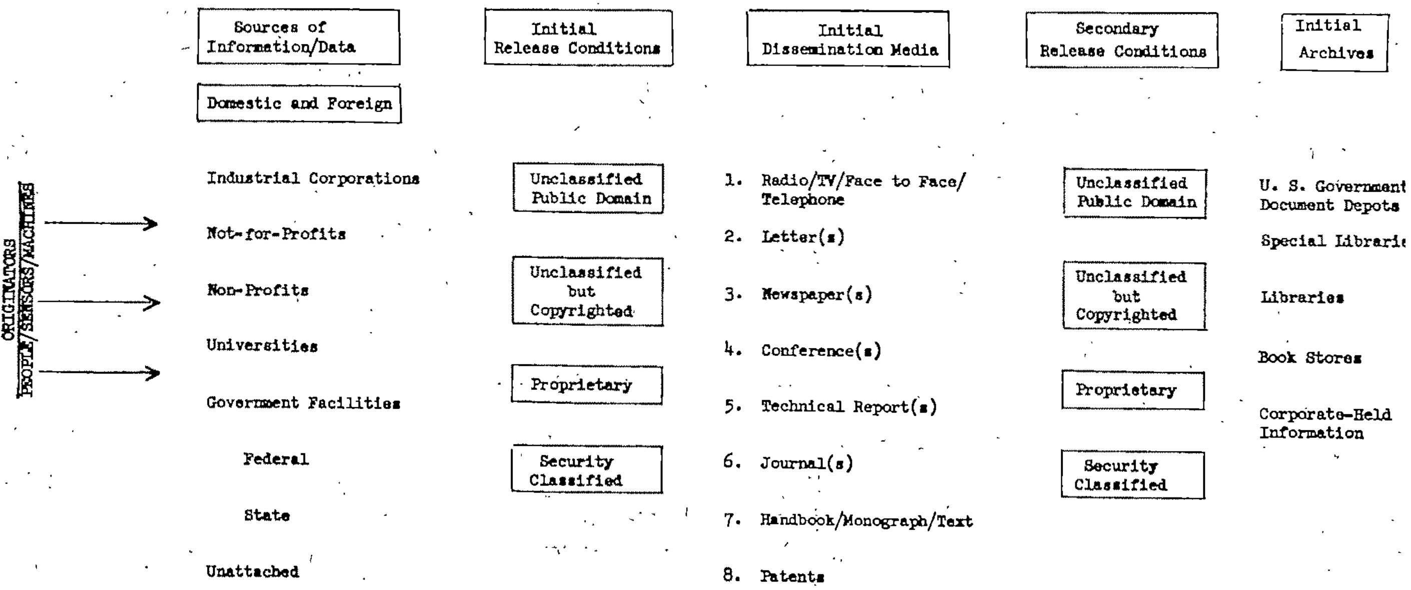


FIG. 1. Information/data transfer spectrum

Handier Writing

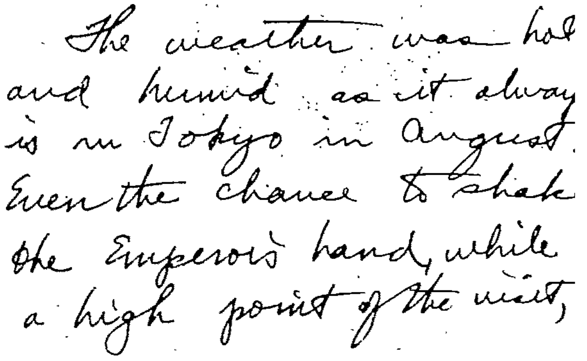
No matter how many machines we have around us, it turns out to be necessary to scribble a note to someone, sometime. And what curious things they are, these notes, with the ever-present danger of being misunderstood, not to mention a kind of contempt for the medium inherent in the commonly poor calligraphy.

Figure 1 shows a sample written in haste and exhibiting most of the possible faults.

Through chance, the opportunity to reform came from a booklet (*1*). Subsequently a (mostly British) collection (*2, 3, 4*) formed itself, and a self-taught version of Chancery script evolved (Fig. 2), a process by no means finished.

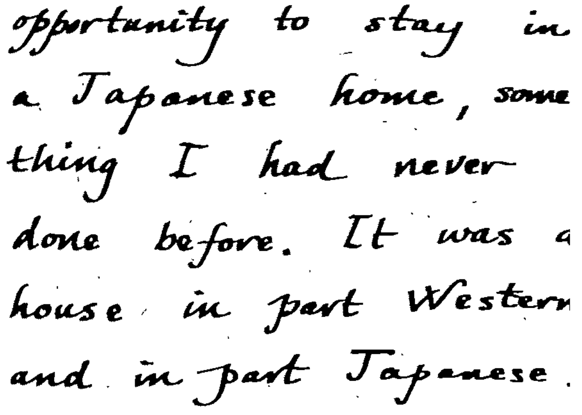
The question most often asked is, "But then my handwriting would look like everyone else's wouldn't it?" Fortunately for self-respect and for cashing checks, individual scripts really are *individual*.

I urge you to try this singular adventure of reforming your script (if it needs it).



The weather was hot
and humid as it alway
is in Tokyo in August.
Even the chance to shak
the Emperor's hand, while
a high point of the visit,

FIG. 1. Before



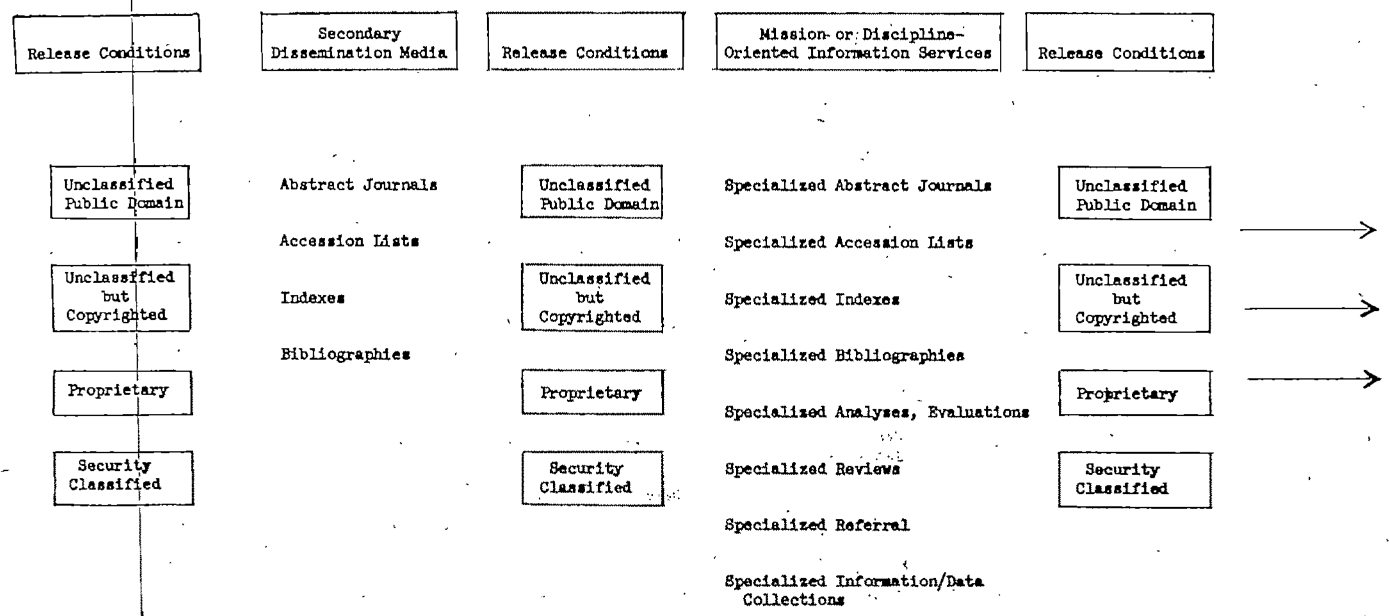
opportunity to stay in
a Japanese home, some
thing I had never
done before. It was a
house in part Western
and in part Japanese.

FIG. 2. After

Selected References

1. TARR, J. C., *Good Handwriting and How to Acquire It*, 3d ed., Phoenix Houise, London, 1954.
2. DUMPLETON, J. L., *Teach Yourself Handwriting*, English Universities Press, London, 1955.
3. FAIRBANK, A., *A Book of Scripts*, rev. ed., Penguin, Harmondsworth, 1952.
4. WEST, A., *Written by Hand*, George Allan and Unwin, London, 1951.

KARL F. HEUMANN
Bethesda, Maryland



Release Conditions
Secondary Dissemination Media
Release Conditions
Mission- or Discipline-Oriented Information Services
Release Conditions
Unclassified Public Domain
Abstract Journals
Unclassified Public Domain
Specialized Abstract Journals
Unclassified Public Domain
Accession Lists
Specialized Accession Lists
Unclassified but Copyrighted
Indexes
Unclassified but Copyrighted
Specialized Indexes
Unclassified but Copyrighted
Bibliographies
Specialized Bibliographies
Proprietary
Proprietary
Proprietary
Specialized Analyses, Evaluations
Security Classified
Specialized Reviews
Security Classified
Security Classified
Specialized Referral
Specialized Information/Data Collections

Letters to the Editor

Dear Sir:

Almost all the world's underdeveloped nations, whether in Asia, Africa, or Latin America, have one common goal: to catch up in every sphere of activity with the developed nations as rapidly as possible. Methods and techniques are of no consequence in their modernization aims.

Recently two items, arriving on the same day in my mail, seemed to gybe. One was a paper delivered by Dr. Roger Revelle, Director of the Scripps Institute of Oceanography, University of California at La Jolla, at the Special National Conference called by the U. S. National Commission for Unesco in New Orleans in September 1966. The paper, entitled "Science and Social Change," urged the utilization of all the social sciences in dealing with both the problems of developed nations (avoidance of nuclear war, big computers and big government, and urban problems), and those of the underdeveloped countries (the population explosion, nutrition and the food supply). To quote Dr. Revelle (*1*):

> Of perhaps even greater importance than scientific agronomy and applied genetics are the economic, social, and political problems of agriculture in the less-developed world. These involve farm credit, marketing, storage, transportation, land tenure, crop diversification, investment in processing agricultural products, and above all, *communication with, and motivation of, the farmers.*

It is perfectly obvious that unless information on the above topics and on technical "know-how" can be made available in very many sections of each agricultural nation in which the need exists either to expand areas of land not presently under cultivation or to multiply crop yields, the possibility of solving the nutrition and food supply problems seems doomed from the start.

The author of the second item I received, Dr. Herman Felstehausen (Assistant Professor of Agricultural Journalism in the Land Tenure Center of the University of Wisconsin, and currently stationed at the Inter-American Land Reform Center, Bogota, Colombia) confirms Dr. Revelle's statement (*2*):

> Availability and distribution of new materials are absolutely necessary if administrators and policy makers are to use the results of research and evaluations to improve development plans and programs. This paper proposes that the agricultural groups in Colombia establish an Agricultural Information Center for the collection and circulation of materials pertaining to agricultural development and the social sciences.

Simply stated, the problem is this: in view of the underdeveloped current status of the typical library (public, university, or special) in the underdeveloped nations of the world, what can be done to utilize this vital communication device in the battle against hunger and malnutrition?

In his paper, Dr. Felstehausen begins by discussing the difficulties in the collecting of published materials of all physical forms in Colombia, following this with data on agricultural libraries in Colombia. He believes the conventional library to be inadequate (in terms of lack of trained librarians, time lags in cataloging, circulation, storage, etc.) and not within the financial means of most Latin American nations.

To meet the needs of agricultural development Dr. Felstehausen presents a plan for an agricultural information library that would utilize perforated cards, cataloging by computer, and various additional nonconventional techniques. Among these would be elimination of the Dewey Decimal (or LC) number in favor of consecutive numbering of items as they arrive. This number would be utilized for circulation, inventory, and shelf location. According to the paper, much of the work in physical preparation would be eliminated; yet, lists of holdings (produced by computers, which are available for rent in Bogota at reasonable rates) could be rapidly produced according to author, by title, subject, date of publication, or publisher. Furthermore, since materials in the field of agriculture consist primarily of pamphlets, folders, and mimeographed papers rather than books, vertical files (or Princeton files on bookshelves) are sufficient.

Dr. Felstehausen goes into the various aspects of his proposal in detail, portraying the obvious advantages of his suggestions over traditional library routines and practices. His final suggestion is the use of a photocopying machine that not only would simplify the library-loan procedure, reduce the amount of items lost, but in addition, for a small charge, would provide the client with his own copy.

Librarianship in Latin America is beginning to display steady progress as witness the feasibility study undertaken by Fundación Interamericana de Bibliotecología Franklin of Buenos Aires, with funds provided by the Rockefeller Foundation and with assistance by R. R. Bowker Company, to survey the possibility of a cooperative cataloging and bibliographic publication center in Latin America (*3*). The recently established publications, *Bibliografía* and *Colbav* of Caracas, Venezuela, and the giant steps taken by librarianship in Mexico and Brazil, not to mention the recent research projects and responsibilities for studies on Latin American library education undertaken by the Inter-American Library School of Medellin, Colombia, are indicators that the time may be ripe for the implementation of Dr. Felstehausen's proposal in the form of a pilot agricultural information center.

What with the personnel (especially cataloger) shortages, the time lags in making items available for circulation, and the availability of electronic equipment at more reasonable rates, it appears that a goodly portion of what Dr. Felstehausen suggests comprises "the handwriting on the wall" for U. S. libraries. All the more reason, then, to apply his proposal to underdeveloped nations that want and need to "leap-frog" into the twentieth century. For many of them it is not merely a matter of prestige, but a question of the survival of masses of their citizens. A forward-looking approach in librarianship and in the imparting of life-giving information in agriculture could conceivably make the difference.

References

1. REVELLE, R., Science and Social Change, *Proceedings of the Special National Conference called by the U. S. National Commission for Unesco, New Orleans, La., 18–20 September 1966*, U. S. National Commission for Unesco, Washington, 1966, p. 37.
2. FELSTEHAUSEN, H., *The Need for Modernized Agricultural Documentation Centers in Latin America: A Colombian Example*, Centro Interamericano de Reforma Agraria, Bogota, October 1966, p. 6.
3. *Noticiero Franklin*, Número 3, Fundación Interamericana de Bibliotecología Franklin, Buenos Aires, September 1965, p. 3.

MARTIN H. SABLE, Chief
Documentation Section
Latin American Center
University of California
Los Angeles, California

Dear Sir:

R. R. Dickison in his brief but informative article on "The Scholar and the Future of Microfilm" in the October 1966 issue of *American Documentation* (pp. 178–179), nevertheless overlooks one factor that will undoubtedly militate against the popularization of the microfiche, as it must have done among the factors enumerated by Dickison with respect to microfilms and microcards. I refer to the importance of making the text of microforms accessible to users through the principal cataloging scheme of the standard collections of the library, using the same classification, the same symbols, tracings, and cataloging rules, and interfiling the resulting entries in the central public catalog of the library, be that in card or book form.

My awareness of the critical importance of this form of organization and processing of the newer unconventional media for their popular acceptance grows out of the repeated frustrations I have experienced in gaining ready access to the pamphlet files and other special collections of a library when these were not accessible through the main public catalog of the library.

To remedy this situation, I have incorporated the subject headings of the pamphlet file in the main public catalog, with gratifying results. The scheme has made the contents of the pamphlet file as readily accessible to users as the holdings in the standard collections of the library in the public catalog.

To the same end I have extended the principal cataloging scheme of the library to the organization and processing of audiovisual materials, in the design of the School Libraries Automation Project (SLAP), an ESEA Title III Operational Grant project to be implemented in the course of three years. The plan also provides for depth access to important, unusual, out-of-the-way materials in all the collections of the library, whatever the size or the form of the media, that are now lost or not readily accessible to users. It does this through a single, coordinated system of indexing and retrieval, without altering the conventional organization of the collections.

The potential of a centralized, unified scheme for the organization and processing of all the collections and types of material, and for gaining depth access to selected items in these collections is so great that the idea has been incorporated by the writer in the design of an automated library system for the worldwide campuses of Friends World Institute, an experimental college in operation since September 1965, and for the Conflict-Resolution Study and Research Center of the Institute to be coordinated with the library resources of the worldwide campuses.

The reason a centralized, coordinated cataloging system for all the collections of a library, other than those covering extremely specialized materials, is important is that the average user tends to limit himself to materials available through the main public catalog, even when he is aware of the existence of other materials in peripheral collections. The material in these peripheral collections may be even more important for his needs than the material in the main collections.

Where these collections are not cataloged as an integral part of the central cataloging scheme of the library, their use may be impaired even for those who may need them most. Their use will be peripheral to the use of the collections accessible through the main public catalog. Important material in the peripheral collections which should nonetheless have priority over the material in the standard conventional collections may either altogether escape the attention of the user, or come to his attention after he has struggled through a maze of secondary materials that have robbed him of precious time that may well have been spent on the more essential sources.

The use of microfilms and microcards would have been many times more since their introduction in libraries, had they been cataloged according to standard procedures and the entries incorporated in the main public catalog. I know of a college that possesses the microcard edition of all the known colonial imprints, but the collection is hardly used because the only way to gain access to it is through Charles Evans' *American Bibliography*. Students as well as faculty may consult it with ease only if they know the date of publication and the name of the author of a work. The subject access to the collection through Evans is clumsy and tedious, and the absence of any entries of the works of the collection in the public catalog of the college library has reduced it to an appendage of little interest or importance among the avalanche of secondary sources that constitute the bulk of the library. What a pathetic waste of financial investment and what an irreparable loss to the academic program of the college to hoard this magnificent, inclusive source of American history, literature, social life, and culture in their varied aspects.

If microform collections are cataloged in the same way as standard collections, and the entries are incorporated in the principal public catalog of the library, their relevance and accessibility in the total program of the library will undoubtedly be multiplied many times over what they are now. This circumstance may have compensated for some of the other factors enumerated by Dickison that have militated against the popularization of the microfilm and microcard. It would assuredly do the same for the microfiche even were its use arrested by the same limitations that affected the use of the microfilm and microcard adversely. Indeed, even if all the limitations listed by Dickison were removed from the microfiche, its use would still be hampered by the relative lack of accessibility to its contents if the cataloging problem raised in this commentary is not properly attended to.

CHARLES A. VERTANES
Sponsor, Library Automation Project
Brentwood Public Schools
Brentwood, New York
and
Director of Research
and Library Consultant
Friends World Institute

Dear Sir:

In the January 1966 issue of *American Documentation*, L. H. Mantell arrives at an estimate of the literary output of scientists and engineers in research and development during 1964 that is very much lower, as he observes, than the estimates that have been made by others. However, Mantell's method has a serious downward bias that vitiates his estimates.

On examining the annual author indexes of a number of journals, Mantell finds the number of authors who have had one publication in each journal, the number who have had two, three, and so on. Assuming that authors were selected for each journal by random sampling with replacement from a population of potential authors (a *distinct* population for each journal), he applies the theory of the Poisson distribution to estimate the total size of the hypothetical population of "authors" from which the actual authors were drawn.

So far so good. Summing over 20 journals, Mantell finds that 2,169 authors have appeared once, 229 twice, 44 three times, 12 four times, and 7 five times and over. The estimate, from Poisson assumptions, of the number of nonpublishing "authors" required to account for these distributions of publications was 9,332 for all the 20 journals.

Next, Mantell smooths these aggregate figures by fitting a Poisson distribution (which fits very well), and thus estimates that 74.1% of the potential authors did not publish at all, 22.2%, published once, 3.3% twice, 0.3% thrice, and 0.1% four times. Applying these rates to an estimated population of 465×10^3 scientists and engineers, he calculates their output at about 140×10^3 titles per annum.

Even if we accept the assumptions from which the Poisson distribution is derived, this estimate is correct only if no "authors" publish in more than one journal. If there is any overlap, Mantell's estimate will be too low. This is easy to show by numerical example. Suppose a population of 100 scientists is publishing in a single journal, with frequencies like those observed by Mantell. Then the total number of papers and the average number per scientist will be just as Mantell estimates them.

Suppose, however, there are *two* journals each drawing its papers independently, by Poisson sampling, from the *same* population of 100 scientists, and with the same total number of papers per journal. The statistics of publication will then be those shown in Table 1.

TABLE 1. Number of authors mentioned, by number of titles per author, in each of two "independent" journals

		Journal B				
	Number of Titles	0	1	2	3	4
Journal A	0	55.0	16.5	2.5	.2	..
	1	16.5	4.9	.7		
	2	2.5	.7	.1		
	3	.2				
	4	..				

We see that 55% of the population will appear in neither journal, 33% will appear once (in one or the other), $2.5 + 2.5 + 4.9 = 9.9\%$ will appear twice (in one or both), and 1.8% will appear thrice. On the other hand, the analogue to Mantell's Table 8 will look like our Table 2.

TABLE 2. Analogue to Mantell's Table 8

	Frequency of contribution per year				
Journal	0	1	2	3	4
A	74	22	3	1	..
B	74	22	3	1	..
Total	148	44	6	2	
Percent	74	22	3	1	

He will therefore conclude that each 100 scientists will contribute a total of $(22 \times 1) + (3 \times 2) + (1 \times 3) = 31$ papers; while from our joint statistics of the two journals, we conclude that each 100 scientists will contribute a total of $(33 \times 1) + (10 \times 2) + (2 \times 3) = 59$ papers, which, except for rounding errors, is just twice the previous estimate, as it should be.

Since we do not know how much overlap exists, in fact, in the potential author pools of different scientific journals, we cannot determine the proper correction to apply to Mantell's estimate. Common observation tells us that the overlap is great and that the errors of estimation must be correspondingly large.

HERBERT A. SIMON
Carnegie Institute of Technology
Pittsburgh, Pennsylvania

Dear Sir:

The Brief Communication by Howard Iker in your January 1967 issue describes a technique for solving Boolean equations that is essentially a programming technique and, therefore, restricted to use within the computer. The computation of the various sums (S) that would represent the "conditions of truth" for a given equation is far too difficult and time-consuming to require of an analyst in a real-life search situation.

I want to make this point because the weighted search described in my own Brief Communication in your July 1966 issue is basically a way to search in its own right and in some respects is easier for the analyst to use and to code than a logical equation. To provide an example: at the present time our SDI system is comprised of over 900 separate profiles, each written as a weighted search. The fact that our weighted search can also duplicate the logic of Boolean equations added to its usefulness. It is not and was not designed, however, as a programming technique for solving equations, but as an analyst's technique for writing search specifications. The basic difference in point of view between the two Communications is, therefore, plain.

As a programming-technique Mr. Iker's system would appear not to be able to handle equations in which a given term makes more than one appearance. For example, see C in the following equation: $(A + B) \cdot (C + D) + (C \cdot E) = \text{ANSWER}$. C cannot be assigned values of both 2^2 and 2^4 simultaneously and it would seem, therefore, that the technique, as described, is limited in the equations to which it can be applied.

W. T. BRANDHORST
Documentation Incorporated
Bethesda, Maryland

Dear Sir:

Mr. Brandhorst's letter raises two issues about my Brief Communication in your January 1967 issue: (1) The method I described is a "programming technique" while that described by Brandhorst is an "analyst's technique"; (2) The method I described will not handle duplicated references to a term in a Boolean equation.

There is no question that the calculation of all possible solutions and the truth-value of each for an equation of any length is indeed a laborious process; indeed, it is a technique designed for a computer, and the reason I mentioned it, as such, was due to Brandhorst's statement that when using the technique he described, "complicated equations can be both difficult and laborious to code. . . . There is no reason that the program should not accept the equation and calculate its own weight assignments. This is now being evaluated." Unless this evaluation has been decided negatively, I still would urge Brandhorst to evaluate the alternate technique I mentioned. This brings us to the second point raised.

Mr. Brandhorst's example of a duplicated term equation (with brackets added for greater clarity assuming the usual priority of "and" over "or") $[(A + B) \cdot (C + D)] + (C \cdot E) = \text{ANSWER}$ represents a set of five elements. Regardless of how many times any equation terms are repeated, a set of five binary elements may be construed in a maximum of 32 ways (including the null-set); there are simply no more combinations available. Given the method I mentioned, $A = 1$, $B = 2$, $C = 4$, $D = 8$, $E = 16$, each of these 32 sets has, by definition, a unique sum; each of these 32 sets may be evaluated for its truth value against the given equation. Accordingly, the equation is false if and only if $S = 0$–4, 8, 12, 16–19, or 24; all other results are true.

The computation of the truth value for each of these 32 sets is indeed tedious, but there seems little question that the process can be done by a computer and equally little question that the implementation of such a set of translated equation weights can be handled efficiently for term-searching.

I hold no brief for the method as the better of the two suggested. What I am suggesting is that it not be ruled out on the grounds of difficulty in coding, which can be handled by a computer, nor of its inability to handle certain kinds of equations since the latter does not seem to be true.

HOWARD P. IKER
School of Medicine and Dentistry
The University of Rochester
Rochester, New York

Book Reviews

2/67-1R **Annual Review of Information Science and Technology.** Volume 1. 1966. Carlos A. Cuadra, Editor. American Documentation Institute. Interscience Publishers, a division of John Wiley & Sons, New York, 389 pp.

This initial volume of the ADI Annual Review calls for kudos to the National Science Foundation, to System Development Corporation, to the American Documentation Institute, and especially to Carlos A. Cuadra for making this needed and useful tool a reality and thus filling one of the gaps in the field of documentation (broadly defined). Critical annual reviews of the literature are essential in any vital field (even those who are aware of the reviewer's bias for *communication* will not quibble with this), so it is with satisfaction that we may greet the fruition of the lengthy efforts of Dr. Cuadra. It also is appropriate that Charles P. Bourne and Pauline Atherton are represented in this first volume, since their early encouragement led Dr. Cuadra to continue his efforts toward this publication.

Now, what do we have? In Dr. Cuadra's words, "a constructive review of topics of current interest to users, designers, and students of information systems and services" (p. 1). This review is provided in 12 chapters, each the responsibility of its individual editor(s). One of the major tasks was the determination of these divisions and selection (and persuasion) of editors to prepare them. They begin, after an introduction, with Chapter 2, professional aspects (Robert S. Taylor) and Chapter 3, information needs and uses (Herbert Menzel). One of the studies reviewed here, John Martyn's "Literature Seaching by Research Scientists" (pp. 58–59), points up the use of such a tool as this, plus the need for interpersonal communication with others working in similar fields. Chapters 4–6 concern technical problems: Chapter 4, content analysis (Phyllis Baxendale), Chapter 5, file organization (W. Douglas Climenson), and Chapter 6, automated language (Robert F. Simmons). Indexing systems (Chapter 7), despite its 110 references, is one of the chapters criticized for omission of some good work. This, I think, is almost inevitable in a work of this type. We should be grateful to its editor, Charles P. Bourne, for his excellent table, Brief Summary of Experimental Evaluation Projects Reported in the Literature (pp. 176–179), and also for his list of methodological questions that "would be good research targets for some of the students in the graduate library schools" (pp. 180–181).

New hardware (Chapter 8, Annual Review Staff) and man-machine communication (Chapter 9, Ruth M. Davis) are followed by three chapters (Chaps. 10–12) on applications: Chapter 10, system applications (Jordan J. Baruch) points up advances in the fields of business, chemistry, drugs, education, law and patents, and the military—the latter naturally limited by what information has been released. Chapter 11, library automation (Donald V. Black and Earl A. Farley), is another area subjected to criticism—e.g., "Why was this paper by Bruce Stewart included and not his other one?"; or, "Why was ISO Planning Memo No. 3 not listed?" etc. There are any number of possible reasons, but this points up a problem of which we need to be aware—that reviews are selective. We cannot sit back comfortably with the book at hand and relax, thinking that now we have all we need. We do not. We do have a great deal more than we had before, and we need to use and support the work to see that this series will be continued. But we need to supplement it with things like *Documentation Abstracts* and other journal literature, and a constant "nose for news"—those interpersonal contacts that are such an important part of communication and awareness in an active, changing field. Here I also want to make a plea for publication. I know of a handful of good projects recently brought to successful completion in the library field which have not been reported on to anyone except the administration. This knowledge and experience should be shared with the profession; it might well bring assistance on some of the problems as well as help others who are contemplating similar investments.

Which brings us also to Chapter 12, information centers (G. S. Simpson, Jr., and C. Flanagan). Of special interest here is the listing of "New Services of 1965" (pp. 318–320). The last chapter, Chapter 13, national issues and trends (John Sherrod), concludes with the statement of belief "that a working plan for setting-up a national information system for science and technology will evolve in the near future" (p. 350).

We have progressed to the keys to specific data in the text—the indexes (Pauline Atherton and Stella Keenan). These 31 pages serve their purpose well, with boldface indicating chapters, and with useful cross-references. The appended Acronym/Abbreviation List is very helpful. My two small quibbles concern the *M's* in the index. I am puzzled to find the *Mc's* at the beginning of the list (before "Ma," "Maass," etc.) in what I understand is a "handmade" list. This does not follow any standard alphabeting practice of which I am aware. The other item is more unfortunate: the name of J. J. Magnino (p. 156) is misspelled to "Mangino, J. J.," and filed as such. However, I think those who know his work would glance at the page, as I did, long enough to spot his initials several lines farther down. (The book as a whole is remarkably free of typographical errors; I found only four others.)

To sum up, we have a fine piece of work here, and a useful one. It is now up to us to see that the work will be continued. Dr. Cuadra is continuing as editor for Volume 2, to cover 1966 (calendar year) literature. We need to write reports on projects, and send a copy to him, as our part in aiding the communication process. We also need to let him hear our suggestions. I have one, which may be shared by others who have worked on United States of America Standards Institute subcommittees: The Institute's recommendation for indexes is that they be in dictionary arrangement in a single alphabet; I suggest that this arrangement be adopted for future indexes. In my opinion, 1965 represents the high-water mark of Dr. Cuadra's contributions to ADI and the profession.

NATALIE C. BATTS
Columbia University Libraries

2/67-2R **Looking Forward in Documentation, Papers and Discussion, Aslib 38th Annual Conference, University of Exeter. 1964.** 1966. Aslib, London. 109 pp.

Whether these 17 papers and the discussions that they stimulated actually are, as the title purports them to be, a perspective of future developments in documentation is at least debatable. To be sure, the contents of this volume do reveal in rather striking fashion the growing interest in and experiences with varying systems of information retrieval as they have developed in Great Britain during recent years, and in this respect the compilation stands out in marked contrast to the typical attitude toward documentation that characterized British documentalists and librarians of only a short time ago. But certainly the work reported can scarcely be regarded as pioneering, and much that is said in these pages will be well known to an American audience.

Of the 17 papers that comprise this collection, 11 are concerned with information storage and retrieval, 2 with library use and user needs, and 4 with primary and secondary publication, including a particularly interesting paper by Sir Thomas Scrivenor on the growth of scientific literature. This distribution of topics probably represents, as Christopher Hanson points out in one of the discussion sessions, the current balance of effort and interest among documentalists in the British Isles.

Our existing store of information is not materially augmented by the papers on information retrieval which, as the above statistics indicate, make up a substantial portion of the entire volume. T. N. Shaw, of the Unilever Research Laboratories describes, for example, a coordinate-indexing system for company reports for the organizing of which he uses a card sorter. His associate at Unilever, H. East, reviews the use of the IBM 1620 computer and the IBM 870 Document Writing System for the processing of published information. Other papers follow the same general expository pattern. Th. W. te Nuyl, from Shell International Research at the Hague, describes the "l'Unite" system using a Texoler Sorter for multiple search. Mannix and Whitehall, also of Unilever, describe a punched-card indexing scheme for organic chemistry, using a fragment code that makes possible the identification of reactions. R. Moss of Shell Chemical of London, considers the problems of vocabulary control when using Batten (Peek-a-Boo) cards and urges the need for further research in that area.

R. C. M. Barnes, of the Atomic Energy Research Establishment at Harwell, argues inconclusively and without visible supporting evidence that computers "can have only a marginal effect upon the quality of retrieval." J. R. Sharp, of British Nylon Spinners Ltd., describes an experimental index—"Slic" for Selective Indexing in Combination —in a paper that has been subsequently published in his text, *Some Fundamentals of Information Retrieval.* J. C. R. Yeates, of the Rowett Research Institute, describes an index created by breaking down titles into their syntactical elements, then reassembling them into a statement that can be used as a basic index entry; additional entries are produced by permutations of this statement. In devising this system, which might be characterized as a slow KWIC index, he has identified 15 types of syntactical elements: Substance, Property, Aspect, Process, Agent, Modification, Modifier, Contrast, Locus, Environment, Condition, Time, Viewpoint, Intention, and Literary Form—all of which makes Ranganathan's famous PMEST formula appear surprisingly simple.

H. J. Zwillenberg, of the Weapons Research Establishment at Salisbury, South Africa, finds that the use of computers for the preparation of bibliographies is economically practicable, even for small-volume information requirements. C. D. Batty, of the Birmingham School of Librarianship, holds that the primary reason for including a consideration of mechanized information systems in the curriculum of the library school is "to demonstrate [the] use [of such systems] to student librarians who will then have relevant experience when later they work with and perhaps introduce similar systems."

The topics considered at the conference which were not directly related to information retrieval were, in the opinion of this reviewer, of considerable interest, especially the one by Sir Thomas Scrivenor, of the Commonwealth Agricultural Bureau. Sir Thomas points to numerous discrepancies in published figures purporting to prove the existence of a "literature explosion" and advances specific suggestions on how best to quantify the true amount of scientific publication. Really valid data, he rightly believes, would "convert Cassandra-like prophecies of impending catastrophe into verifiable factual statements." Barnes, at Harwell, finds that inquiries suitable for computer processing account for only a small proportion of the total received by his information office, "perhaps only nine per cent and certainly not more than thirty-five per cent." Furthermore, delays resulting from the time required for the preparation of a search plan and the computer search itself "would result in a service appreciably slower than that provided by present methods." John Martyn and Mrs. Margaret Slater, of Aslib's research department, present some very interesting, though tentative, conclusions concerning the behavior characteristics of users of scientific information in libraries; their study should be substantially extended. Gordon Y. Craig of the University of Edinburgh, urges publishers of scholarly journals to initiate the issuing of abstract cards for the articles in their periodicals. F. Liebesny, of British Aluminum Ltd., presents *Aluminum Abstracts* as a laudable example of international cooperation in the abstracting field. M. E. L. Morris, of University Microfilms Ltd., writing under the intriguing and not too permutable title of "From One to Two-Fifty," proposes the formation of an international agency to collect and package requests for microphotographic copies of wanted materials. Such packaging would, he believes, substantially reduce the costs of microfilm requisition. Phyllis I. Edwards, of the Department of Biology, British Museum, believes, after a survey of users of abstracting services in the biological sciences, that a British association of science abstracting and indexing services, comparable to that now active in the United States, is badly needed to coordinate and otherwise improve the quality and coverage of such facilities.

Unfortunately, the most interesting papers of all are missing from this collection; these papers might have reported the tentative results of research in progress at Aslib. But the investigations are listed in the present volume only by title with a brief note on the aim, method, and progress to date of each. One would like to know much more about such investigations as: a case study in depth of the information needs of scientists; the demands of users of technical libraries; economic factors in I. R. systems; comparative tests of small and large systems in the same subject field; and the fabrication of model indexes. All these topics are of extreme interest to American librarians and documentalists, and their published results will be eagerly awaited.

If the reviewer may be pardoned a nationalistic note, he would like to point out that, though the book under review is not heavily documented, of the total of 42 citations it records, exactly half (21) are from American sources. These remaining are divided among England (11) and other countries (10). Such statistics may have no meaning whatever, though they might seem to suggest that the British are at least keeping a watchful eye on the documentation efforts of their American cousins. But whatever conclusions one can or cannot draw from such a count, the fact remains that the librarians and documentalists on John Bull's Island have made substantial progress in the past decade in developing unconventional approaches to the subject analysis of library and bibliographic materials. We can well remember the conference sponsored by Aslib at Dorking in 1957, when the group, by fiat, declared that any mention of computers and mechanization *must* be limited to "Thursday morning when the Americans can have their say." The mention of machines at any other time was strictly *verboten.* One cannot but wonder whether the slow entry of the British into mechanized retrieval activity has been due to canny native caution or inadequate financial resources.

J. H. Shera
School of Library Science
Western Reserve University

2/67-3R **Information Management in Engineering Education.** (Proceedings and recommendations of the Conference on Information Sources, Systems and Media in Engineering Education, held at Lehigh University on May 19–20, 1966). 1966. Robert S. Taylor, Editor. Lehigh University, Bethlehem, Pa. 76 pp.

Surely no one is apt to criticize the goal of this conference, for the object was to develop a national plan for improving the utilization of technical information by engineers. The main objective was to produce better courses for engineering students in which they would learn to appreciate and evaluate the many facets of such information. However, not everyone will agree with the general makeup of the body of participants at the conference or with all the recommendations in the plan they produced.

The conference was jointly sponsored by the American Society for Engineering Education and by Lehigh University. The report lists only 23 participants. Aside from the editor and one or two others, the participants seemed to be either professors of engineering or from technical information departments (such as computer centers and the like) in industry or universities. The term "information management" was coined by one of the participants to signify the understanding of and utilization of information sources, media, and systems. The participants obviously were not content merely to sigh over the problem of better utilization of information, for they came up with a very detailed plan of action that included at least three pages of flow charts and numerous lists and categories. In fact, it was this very amount of detail that made this report so unlike the usual conference proceedings. Aside from the account of the introductory talk by the conference chairman, the remainder of the report reads like the final document of a study commision, with summaries and a thorough discussion of all the aspects of a problem. There is also a two-page list of references, which was given the rather unusual listing in the table of contents as a "suggestive bibliography." One fault with the references was that so many were to unpublished literature. But, in general, one admires the goal of the group and their diligence in preparing such a comprehensive plan.

There was a general summary plus the reports and recommendations of four panels. The introductory talk by the chairman gave background information on the relationship of engineers to information utilization and also listed several questions for the conference to answer. The panel on Objectives described the outline of their national plan, as well as discussing proposed courses for the new project. The panel on Curriculum elaborated on the contents of the courses, including a comparison of the "problem" method of presentation versus the "formal" method. The panel on Logistics and Implementation gave more of the details of the national plan, including summer training institutes for the instructors of the courses, setting up task forces for planning the curriculum, etc. The panel on Support and Evaluation dealt with the establishment of a permanent secretariat, with funding (at least $1,300,000 for a three-year program) and various evaluation techniques.

Regardless of the outcome of this particular program, it is at least encouraging to see such interest by engineers and engineering school faculties in improving these conditions. However, one aspect of the conference that seemed to me to be regrettable was th apparent lack of participation by those now engaged in the teaching of courses in this subject for engineering students. Courses have been given for many years at schools all over the U. S. Did the conference have the benefit of the experience of any of these instructors, or even of a spokesman for the group? If so, it is not apparent from the list of participants. Surely some real benefits could have been derived from hearing about their successes and failures in actual classroom work.

Another point that bothered me was the recommendations for instructors for the new courses. It appears that engineers on the regular engineering school faculties will be the chosen ones. For this reason, one part of the plan was devoted, at considerable length, to summer training institutes for preparing the engineers for the teaching of the new courses. This seems again to overlook the contribution that could be made by those now engaged in such work. Granted that the existing courses are not perfect, it would seem to me most likely that present instructors, if they have any competence at all, would be pleased to work towards improving the course content and methods of presentation, and that this upgrading could be done more quickly and more easily than starting more or less from scratch with the recruits to go to summer training institutes. Almost any engineering school worth its salt should have technical librarians and documentalists and data processing managers who, singly or jointly, could do a very creditable job of teaching such courses. Many of these people have had collegiate training, or even degrees, in the sciences and engineering as well as training in their library or documentation duties. It is not unusual to find librarians who are familiar with all the latest trends in using computers, or documentalists who are familiar with the reference tools used in literature searching, etc. On the other hand, engineers, unjustly or not, do not enjoy a good reputation as being very familiar with technical information resources in general and as a class are not usually rated as being "heavy" users of this information. Why not use the people whose job it is to organize, select, use, and manipulate technical information on a daily basis to serve as the instructors of the improved courses?

This report may very well be followed by a large-scale improvement in the utilization of technical information. The improvement is long overdue. But let us hope that it is done with the active participation of those who are experienced already in such matters. Librarians and documentalists have been concerned about this problem for many, many years. Why not put them on the team too?

Ellis Mount
Science & Engineering Librarian
Columbia University

2/67–4R **Education and Training of Information Specialists in the U. S. A.** May 1966. Marilyn C. Bracken and Charles W. Shilling. Biological Communication Project, Washington, D. C. 70 pp.

This is a brief survey of the current status of information education, which proves to be a larger and more varied field than most readers will realize. The field is interpreted broadly to include some computer instruction, often administered centrally in the computer center, as well as some traditional library science course work. Fifteen doctoral programs are included, but the possibility of getting a strong concentration in information science is limited to only a few of them, most of them offering little more than library science. There is a list of the ALA-approved library schools with an indication of the information science courses offered. Another section of this paperbound booklet describes the programs of 20 leading schools in two or three pages apiece. The faculty lists are sometimes helpful but include many persons only remotely concerned with information science education. Another table lists professional associations and their interest in information science education. A bibliography of recent information concludes the booklet.

A few of the items of information are either incorrect or misleading, but in general this is a useful compilation that will go out of date quickly. While it is current, those for whom superficial factual data on the field is important will need it in their offices.

John Harvey, *Dean*
Graduate School of Library Science
Drexel Institute of Technology

2/67–5R **A Plan for Indexing the Periodical Literature of Nursing. Report of a Study of the Need for Bibliographic Control of the Scholarly Record of Nursing.** 1966. Vern N. Pings. American Nurses Foundation, New York. 202 pp.

The serial literature of nursing is indexed by two organizations. This monograph recommends that one of these two continue indexing such literature, and that other organizations interested in the literature of nursing cooperate with it in disseminating the results. A plan for the National Library of Medicine routinely to index the nursing literature found in serial titles coverd by *Index Medicus* is succinctly stated in part of the book. The rest of the book is devoted to background on the history and development (or lack thereof) of nursing librarianship and bibliographic control, accompanied by statements on both nursing education and the profession.

In actual fact, the plan presented is already in operation. In the book, it would be helpful if the justification of the need for the particular index plan recommended were as strongly presented as is the conclusion that the index is needed and that this particular plan for accomplishing it through the facilities of the National Library of Medicine's generalized system is the means that best serves those interested in the nursing literature. Comparisons of an analysis of the products of the *Cumulative Index to Nursing Literature* with that of the National Library of Medicine are made. Not sufficiently emphasized is the importance of NLM's routinely screening the journals covered by *Index Medicus* over the fact that NLM uses a computer in manipulating the citations. Only a beginning is made on enumerating the inherent and administrative drawbacks of the operational computer system used. There was a technical oversight in calling GRACE (Graphic Arts Composing Equipment) graphic arts company equipment, which illustrates that the NLM system was perhaps not observed closely enough.

The monograph is strongest when presenting background information on nursing librarianship and bibliographic control, and in its many tables of data illustrating the quantification and categorization of data relating to the subjects covered in the various chapters, titled as follows:

1. The Quest for Quality in Nursing
2. A Review of the Efforts of Professional Groups to Achieve Bibliographic Control of the Literature of Nursing, 1900–1964.
3. Analysis of the Coverage of Nursing Literature in the Medical Literature Analysis and Retrieval System
4. Comparison of Indexing in *Cumulative Index to Nursing Literature* with Indexing in MEDLARS
5. Quantitative Characteristics of Journal Literature of Interest to Nurses
6. The Possible Users and Purchasers of an Index to the Periodical Literature of Nursing
7. Cost of Publishing an Index to the Journal Literature of Nursing Utilizing MEDLARS Facilities
8. Possible Administrative Relationships among Agencies now Concerned with the Bibliographic Control of the Journal Literature of Nursing
9. A Review of the Published Literature on Nursing Libraries, 1900–1963
10. Access to the Scholarly Record of Nursing
11. Study of a Plan for Indexing the Periodical Literature of Nursing

The merit in and the importance of the book lie in its example of an approach to analyzing and evaluating information systems, and in its revealing the condition of nursing bibliographic control complete with extensive references. These alone more than justify its purchase and perusal. In addition, it records the background that led to the cooperative effort between the National Library of Medicine and the American Journal of Nursing Co. in producing the *International Nursing Index*, which may set the pattern for similar ventures in other medical and paramedical fields where NLM is indexing literature and the specialty field concerned is financially hard pressed to publish and index the literature for its area.

HAYDEE GARCIA CLARK
Librarian, Chestnut Lodge
Rockville, Maryland

2/67–6R **Indexing and Classification; A Selected and Annotated Bibliography.** A joint project of the Nuclear Science Division and Documentation Division of the Special Libraries Association. May 14, 1966. Compiled and edited by Winifred F. Desmond and Lester A. Barrer. Oak Ridge National Laboratory Library, Oak Ridge, Tennessee. 256 pp. (on microfiche)

Two divisions of the Special Libraries Association have here joined together to produce the first SLA publication in microfiche form. The 356 typescript pages of this annotated bibliography of recent publications in the fields of indexing and classification are contained on six 4 × 6 microfiches.

The 635 entries of this bibliography are intended to cover the field fairly comprehensively. Indexing and classification are interpreted quite broadly to include general books on data processing and information retrieval. Citations are given for technical reports, monographs, periodicals, conference proceedings, etc.

But the scope is somewhat confusing. The list is called a selective bibliography because of the omission of classification schemes, glossaries, and vocabularies. But a few outstanding examples of this latter group *are* included as samples. The time period covered is January 1, 1960, to mid-1964. But a few publications of the 1950's *are* included because of their special significance. The citations are essentially to English-language publications. But a few items in other languages *are* included. With all of these exceptions, it is probably worthwhile to look for almost any important publication on indexing of the last 15 years or so in this bibliography, even if it falls outside the formal limitations.

The entries are arranged alphabetically within a broad classified order. Reports are entered under the organization responsible for the research. Journal articles and monographs are entered under personal author, corporate author, or title. Papers in proceedings are also entered under author.

Information is supplied about the availability of technical reports. When a report is available through the Clearinghouse for Federal Scientific and Technical Information, the price is given.

The annotations for each entry come from *American Documentation, Computing Reviews, Computer Abstracts,* and a few other journals. When no annotation was available in the literature, the editors provided one of their own. Beneath each abstract are listed the indexing terms assigned to the document.

Some interesting statistics on the nature of indexing literature are provided in the Introduction. A frequency count was made for the types of sources of entry included in the *Bibliography*. Journal articles led with 229. Conference literature followed with 146; report literature was close behind with 125. Of the 51 journals that contributed entries, *American Documentation* led with 43 articles. The *Journal of Chemical Documentation* followed with 38. No other periodical source came even close.

The most potentially interesting (and ultimately disappointing) feature of this bibliography is its set of indexes. Besides a computer-produced author index (enriched manually when authors were not main entries), two subject indexes are provided. One is a computer-produced KWIC permuted title index; the other is a manually-produced subject index. The Preface describes this as an opportunity to make a comparison between computer KWIC indexing and manual keyword indexing.

It is not clear what sort of comparison was really intended. Manual techniques can achieve more sophisticated indexing than is now possible with a computer. The only valid points of comparison would seem to be relative times and costs for similar levels of indexing. But no such figures are provided.

What is more disappointing is that neither index is really adequate. The KWIC index falls into many of the expected traps. Words like "indexing," "information," and "retrieval" appear in great abundance, filling up pages without being much help to the user of a bibliography on indexing and information retrieval. There is no vocabulary control, which means there is much filing under synonyms and under different forms of the same word, e.g., "index," "indexes," and "indexing." The only aid given the user is the context of the title in which the term appeared.

The manual index would have served the purposes of comparison if it had simply shown how far our machine techniques must still be developed before they can achieve the quality of good manual indexing. But what we have here is not good manual indexing. The subject terms are drawn from the titles *and the abstracts*. Thus, we *are* given a number of subject entries missing from the permuted title index. But we again find entries like "indexing" and "automatic information processing" followed by long lists of indentifying document numbers. There isn't even any context here to serve as some sort of aid in differentiation.

In the case of the KWIC index, we can say that the machine does not know any better. But there is no excuse for including the term "indexing" in a manual index devoted to indexing.

The Special Libraries Association has provided us with a valuable bibliography, and has then failed to index it properly (even though it tried twice). This is especially unfortunate for a publication on microfiche in which skimming and flipping back and forth is somewhat more difficult than in an ordinary book.

ALAN R. GREENGRASS
School of Library Service
Columbia University

Publications of the

AMERICAN DOCUMENTATION INSTITUTE

AMERICAN DOCUMENTATION,* quarterly journal of the ADI. Subscription rate: $18.50 per year, plus $.50 postage for foreign subscriptions. (Subscription included in annual membership dues of $20.00.)

EDUCATION FOR INFORMATION SCIENCE,† proceedings of a symposium held September 1965. $2.00 to ADI members, $6.00 to nonmembers.

Papers and Proceedings of ADI Annual Meetings—

AUTOMATION AND SCIENTIFIC COMMUNICATION ‡ (1963). $9.50 to ADI members, $12.50 to nonmembers.

Beginning with the 1964 volume, the ADI Annual Proceedings are in a numbered series—

PARAMETERS OF INFORMATION SCIENCE † (Volume I, 1964). $7.85 to ADI members, $15.75 to nonmembers.

PROCEEDINGS OF THE 1965 FID CONGRESS † (Volume II, 1965). $9.30 to ADI members, $10.95 to nonmembers.

PROGRESS IN INFORMATION SCIENCE AND TECHNOLOGY § (Volume III, 1966). $12.00 to ADI members, $16.00 to nonmembers.

AND TWO NEW PUBLICATION VENTURES OF ADI—

DOCUMENTATION ABSTRACTS,‖ a quarterly abstract journal designed to be a comprehensive source of information about the literature of documentation and related areas—initiated in 1966 and published jointly by the American Documentation Institute, the Chemical Literature Division of the American Chemical Society, and the Special Libraries Association. 1967 subscription: $15.00 to ADI members, $25.00 to nonmembers.

ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY,♯ a new series devoted to consolidating the latest developments of the growing field of information science and technology. The series will not merely reflect or cater to current interests; it will attempt, also, to broaden and deepen them. Volume I, 1966, is $12.50 to nonmembers of ADI, and $10.63 to ADI members *if members order from the American Documentation Institute.*

*Order from

AMERICAN DOCUMENTATION INSTITUTE
2000 P Street, Northwest
WASHINGTON, D. C. 20036

PAYMENT WITH YOUR ORDER IS REQUESTED

Other titles should be ordered as follows:

† Spartan Books, 1250 Connecticut Avenue, N.W., Washington, D. C.
‡ Kraus Reprint Corporation, 16 East 46th Street, New York, New York
§ Adrianne Press, P.O. Box 644, Woodland Hills, California
‖ Documentation Abstracts, Inc., P.O. Box 9018, Southeast Station, Washington, D. C. 20003
♯ John Wiley & Sons, 605 Third Avenue, New York, New York

ICAN DOCUMENTATION INSTITUTE 1967 ANNUAL CONVENTIO

CALL FOR PAPERS

The 1967 Annual Convention of the American Documentation Institute will be held in New York City at the Hilton Hotel, October 22-26, 1967. The convention's theme is

LEVELS OF INTERACTION BETWEEN MAN AND INFORMATION

We invite you to present a short paper which will amplify our theme by reporting on related and significant techniques, trends and achievements. Papers should, if possible, conform to one of these outlined topics:

THE CREATOR OF INFORMATION
. . . the creative writer, the scientist, the graphic artist, the editor and the publisher.

THE USER OF INFORMATION
. . . using information in the business world, man/machine interface, psychology and information, language and information.

THE HANDLER AND PACKAGER OF INFORMATION
. . . traditional methods of organizing and storing information, new approaches to organizing information, wares of information services (indexes, catalogs), information sharing -- the emerging network.

The short paper should not exceed 2000 words. Text, illustrations, bibliography, etc. must fit on five 8-1/2 x 11" typewritten pages. Detailed instructions for camera-ready copy will be sent on receipt of the attached reply form.

Authors will be grouped by topic for panel discussions and allowed to give ten-minute precis of their papers. All selected papers will be printed and distributed at the convention.

DEADLINES: JUNE 1, 1967
The Program Chairman should be notified by June 1, 1967 of your intent to submit a paper. Use attached form.
JULY 1, 1967
Short papers should be submitted to the program chairman by July 1, 1967.
AUGUST 1, 1967
Authors of selected short papers will be notified of selection by August 1.

SEND REPLY TO: Paul Fasana, Program Chairman
ADI 1967 Annual Convention
Columbia University, The Libraries
New York, N.Y. 10027

REPLY FORM - ADI 1967 CONVENTION

To: Paul Fasana
The Libraries
Columbia Univ.
New York NY 10027

I plan to submit a short paper in the following subject area ______________
____

working title ______________________________________
____

Name ______________________________

Affiliation ______________________________
____

Mailing address ______________________________
____
____

Telephone

Consolidates the latest developments in the growing field of inquiry concerned with information.

Annual Review of Information Science and Technology

American Documentation Institute

Volume 1

Edited by CARLOS A. CUADRA,
System Development Corporation.

This series is intended to be of tangible benefit to individuals, public and private organizations, universities, and government agencies. Although the field cuts across numerous disciplines, the people involved in it all share a concern with the generation, transformation, communication, storage, retrieval, and use of information. This volume, and the ones which will follow it, are planned as constructive reviews of topics of interest to users, designers, and students of information systems and services. By providing a perspective on the information field, such reviews not only point out gaps and duplicative work but also serve as sources of new ideas.

A long-range goal for the *ADI Annual Review* is to encompass the larger communication process in which documentation plays such an important role. To this end, the authors have made an effort to examine not only the obvious literature but also that in psychology, sociology, communication, engineering, management, business, and other fields that have a significant bearing on the communication process. *The Annual Review* will attempt not merely to reflect current interests; it will also attempt to broaden and deepen them.

ADI's *Annual Review* will be keyed to the calendar year. The present volume covers, for the most part, literature that appeared in the calendar year 1965. Some of the earlier literature is also reviewed because this volume, as the first in the series, had no prior context to serve as a frame of reference. Volume II, which will appear in the fall of 1967, will cover 1966 literature. An Interscience Series. *1966 389 pages $12.50*

A discount of 15% is available to members of the ADI if volumes are ordered through the Institute.

of related interest . . .

National Document Handling Systems for Science and Technology

By LAUNOR F. CARTER, GORDON CANTLEY, JOHN T. ROWELL, LOUISE SCHULTZ, HERBERT R. SEIDEN, EVERETT WALLACE, RICHARD WATSON, and RONALD E. WYLLYS, *all of the System Development Corp., Santa Monica, California.*

This new book is the result of the COSATI- (Committee on Scientific and Technical Information) sponsored study on national information systems relating to scientific and technical documents. Presented in a clear and well organized manner, the emphasis of the study, as stated by COSATI, is—

1. Initial and primary priority will be placed on national systems relating to scientific and technical documents, their handling and the management of such documents.
2. Secondary attention will be given to development of programs which can be undertaken with government support for identifying, analyzing, and giving a structure to the total flow of scientific and technical information in this country.

1967 344 pages $9.95

The Analysis of Information Systems

A Programmer's Introduction to Information Retrieval

By CHARLES T. MEADOW, *IBM Corporation.*

Treats information retrieval as a communication activity among a user, a library, and an author. The author assumes some computer background. *1967. 301 pages. $11.50.*

Growth of Knowledge

Readings on Organization and Retrieval of Information

Edited by M. KOCHEN,
The University of Michigan.

A selection of essays on information retrieval, stressing the importance of evaluating and synthesizing newly generated knowledge into a coherent overall image. (*A volume in the Library of Behavior Science Series*). *1967. Approx. 368 pages. $12.95*

John Wiley & Sons, Inc.
605 Third Avenue, New York, N. Y. 10016

ten unique information services

1. **ASCA® (Automatic Subject Citation Alert)**—our computer searches the literature as fast as it appears and alerts you each week to specific items relevant to your interests.

2. **Science Citation Index®**—for the period indexed, tells what works cite specific earlier works, providing retrospective searching. Published quarterly; cumulated annually. Available for 1967, 1966, 1965, 1964 and 1961.

3., 4., 5. **Current Contents®**—your weekly guides to what's appearing in more than 1,600 domestic and foreign journals. Published in three editions: **Physical Sciences, Life Sciences** and **Chemical Sciences.**

6. **Index Chemicus** —weekly graphic abstracting and indexing service for researchers who need fast, accurate and thorough reports about new chemical compounds and their syntheses.

7. **Encyclopaedia Chimica Internationalis™**—cumulates **Index Chemicus™** yearly with specialized rapid-search indexes for retrospective searching. Volumes for 1966, 1965, 1964, 1963, 1962-63, 1960-62 available separately or as complete 23-volume set.

8. **ISI Search Service**—when information problems hold up your work, personalized searches by **ISI** information scientists bring fast, pertinent answers.

9. **ISI Magnetic Tapes**—delivered weekly for use in your own information system to search the most comprehensive literature file available anywhere.

10. **OATS™ (Original Article Tear-Sheets)**—one-day delivery of the original journal pages of any article reported, abstracted or indexed by any **ISI** services.

Please send me details on specific ISI SERVICES:

- □ *ASCA (Automatic Subject Citation Alert)*
- □ *Science Citation Index*
- □ *Current Contents Physical Sciences*
- □ *Current Contents Life Sciences*
- □ *Current Contents Chemical Sciences*
- □ *Index Chemicus*
- □ *Encyclopaedia Chimica Internationalis*
- □ *ISI Search Service*
- □ *ISI Magnetic Tapes*
- □ *OATS (Original Article Tear Sheets)*

name ____________________

title ____________________

organization ____________________

address ____________________

city ____________ state ____________

zip ________ country ____________________

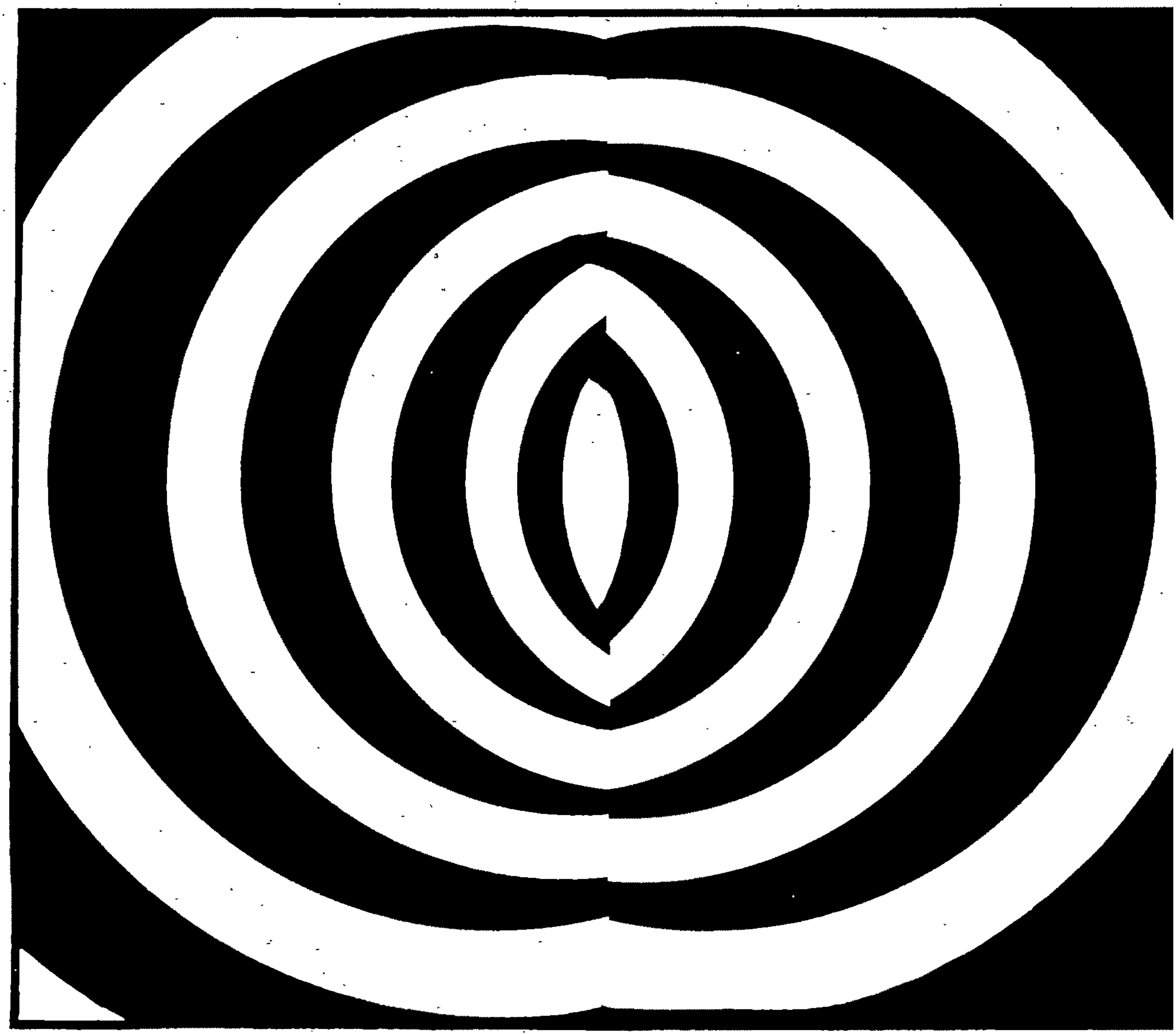
07-6

INSTITUTE FOR SCIENTIFIC INFORMATION • 325 Chestnut St Philadelphia Pa 19106 USA

nerican Documentation

JULY
1967
Vol. 18, No. 3

PUBLISHED QUARTERLY BY THE AMERICAN DOCUMENTATION INSTITUTE



AMERICAN DOCUMENTATION

INSTRUCTIONS TO AUTHORS

American Documentation is a publication of the American Documentation Institute. It is a scholarly journal in the various fields in documentation and serves as a forum for discussion and experimentation. Papers already published or in press elsewhere are not acceptable. For each proposed contribution, one original and two copies (in English only) should be mailed to Mr. Arthur W. Elias, Editor, *American Documentation,* Institute for Scientific Information, 325 Chestnut St., Philadelphia, Pennsylvania 19106. The manuscript should be mailed *flat* in a suitable-sized envelope. Graphic materials should be submitted with suitable cardboard backing.

Types of Manuscripts: Three types of contributions are considered for publication: full-length articles, brief communications of 1,000 words or less, and letters to the editor. Letters and brief communications can generally be published sooner than full-length manuscripts. Books, monographs, and reports are accepted for critical review. Two copies should be addressed to the Review Editor, Dr. T. Hines, 54 North Drive, East Brunswick, New Jersey.

Processing: Acknowledgment will be made of receipt of all manuscripts. *American Documentation* employs a reviewing procedure in which all mansucripts are sent to two referees for comment. When both referees have replied, copies of their comments are sent to authors with the Editor's decision as to acceptability. The refereeing procedure requires about 30 days. Authors receive galley proofs with a five-day allowance for corrections. Standard proofreading marks should be employed. Reprint order forms are forwarded with galleys.

Format: All contributions should be typewritten on white bond paper on one side only, leaving about 1.25 inches (or 3 cm) of space around all margins of standard, letter-size (8.5 × 11 inch) paper. Double spacing must be used throughout, including the title page, tables, legends, and references. The first page of the manuscript should carry both the first and last names of all authors, the institutions or organizations with which the authors are affiliated, and notation as to which author should receive the galleys for proofreading. All succeeding pages should carry the last name of the first author in the upper right-hand corner (0.5 inch from the top) and the number of the page.

Style: In general, style should follow the forms given in the Style Manual for Biological Journals (SMBJ), published for the Conference of Biological Editors by the American Institute of Biological Sciences (1964).

Title: The title should be as brief, specific, and descriptive as possible. Vague and unrevealing titles may delay publication.

Abstract: An informative abstract of 200 words or less must be included, typed with double spacing on a separate sheet. This abstract should present the scope of the work, methods, results, and conclusions.

Acknowledgments: Financial support may be listed as a footnote to the title. Credit for materials and technical assistance or advice may be cited in a section headed "Acknowledgments," which should appear at the end of the text. General use of footnotes in the text should be avoided.

Graphic materials: *American Documentation* requires finished artwork. Follow the style in current issues for layout and type faces in tables and figures. A table or figure should be constructed so as to be completely intelligible without further reference to the text. Lengthy tabulations of essentially similar data should be avoided.

Figures should be lettered in black India ink. Charts drawn in India ink should be so executed throughout, with no typewritten material included. Letters and numbers appearing in figures should be distinct and large enough so that no character will be less than 2 mm high after reduction. A line 0.4 mm wide reproduces satisfactorily when reduced by one-half. Graphs, charts, and photographs should be given consecutive figure numbers as they will appear in the text; however, figure numbers and legends should not appear as part of the figure, but should be typed double spaced on a separate sheet of paper. Each figure should be marked *lightly* on the back with the figure number, author's name, complete address, and shortened title of the paper.

For figures, the originals with two clearly legible reproductions (to be sent to referees) should accompany the manuscript. In the case of photographs, three glossy prints are required, preferably 8 × 10 inches.

Organization: In general, papers should state the background and purpose of the study, followed by details of methods, materials, procedures, and equipment. Findings, discussion, and conclusions should appear in that order. Appendixes may be employed where appropriate for extensive lists, statistics, and other supporting data.

Bibliography: Accuracy and adequacy of the references are the responsibility of the author. Therefore, literature cited should be checked carefully with the original publications. References to personal letters, abstracts of verbal reports, and other unedited material may be included. If an as-yet-unpublished paper would be helpful in the evaluation of a manuscript, it is advisable to make a copy of it available to the Editor. When a manuscript is one of a series of papers, the preceding member of the series should be included in literature cited.

Citation Format:

Order: Literature cited should be sequentially numbered as cited.

Authors: Give all authors with arrangement as follows:
Elias, A. W., B. H. Weil, and I. D. Welt

Titles: Give full titles of articles in English, indicating language of original as: (In Ger.)

Journals: Journal titles should be given in full.

Monograph and Serial Data: Should be presented in order as follows: Volume, issue number, pagination, and year. The issue number should be given in parentheses if journal pagination is not continuous from issue to issue. Pagination should be inclusive. Year of publication should be given in parentheses. An example is given below:

Bishop, D., A. L. Milner, and F. W. Roper, Publication Patterns of Scientific Serials, American Documentation, 16 (No. 2): 113–21 (1965).

American Documentation is published in January, April, July, and October. One copy is included in the individual membership fee ($20.00 per year), three copies in the contributing membership fee ($100.00 per year), and up to five copies in the sustaining membership fee ($500.00 per year). Nonmembers may subscribe at $18.50 per year, postpaid in the U.S. Single copies may be purchased for $4.65 each. Communications concerning memberships, subscriptions, reprints, renewals, back issues, advertising, and changes of address should be sent to the American Documentation Institute, 2000 P Street, NW, Washington, D. C. 20036.

American Documentation is indexed in *Library Literature, Current Contents of Space, Electronic & Physical Sciences, Library Science Abstracts, Science Citation Index, Chemical Abstracts,* and *Documentation Abstracts.*

American Documentation is entered for second class mailing at Baltimore, Maryland.

American Documentation

PUBLISHED QUARTERLY BY THE AMERICAN DOCUMENTATION INSTITUTE

Vol. 18, No. 3 JULY 1967

Published quarterly by AMERICAN DOCUMENTATION INSTITUTE and printed by Port City Press, Inc., 1323 Greenwood Road, Baltimore, Maryland 21208. Second class postage paid at Baltimore, Maryland. Copyright © 1967 by AMERICAN DOCUMENTATION INSTITUTE.

Editorial

Three new projects are scheduled for American Documentation for 1967 and 1968. All are designed to keep ADI members and AD subscribers in the closest possible contact with our rapidly expanding field while improving the quality of the journal.

The first project was suggested by Dr. Eugene Garfield and represents a logical extension of our refereeing procedure, coupled with the use of these pages as a forum for discussion and review of ideas. Up till now such discussion has been limited in length to that appropriate for our *Letters* and *Brief Communication* sections. Now, the Editor actively solicits your contribution of *Opinion Papers*. These contributions will have the same treatment and format as regular articles, but will allow for extensive discussion of an *opinionated* type. The first of these is scheduled for the October issue and, we hope, will set the pace for a succession of analytical treatments, challenging and provocative in nature.

Continuing this concept, the second project was suggested by the Central Ohio Chapter in connection with their host responsibility for the 1968 Annual Meeting. Authors of all papers published in 1968 (Issues 1–3) will be invited to a special Author Forum at the Annual Meeting. At this forum the published papers will be laid open to further examination, question, and discussion, a procedure designed to be of great interest to authors and Editor alike.

Finally, the Editor has already received and earnestly solicits papers on Copyright and Documentation. These are to be published either as a separate issue or as a continuing series for AD in 1967 and 1968. All viewpoints are welcomed so that this increasingly important topic may be fully examined by the AD readers.

A. W. ELIAS

Watson Davis (1896–1967)

We regret to record the passing, on June 27, 1967, of Mr. Watson Davis. Mr. Davis was the founder of the American Documentation Institute and its President from 1937 to 1947. As scientist, journalist, and documentalist his honors and awards are too numerous to mention. We mourn the loss of a pioneer of our profession.

A.W.E.

The point is that libraries are collections of files, files that have to be updated and manipulated so that you can search them from various angles. No one can challenge the fact that file manipulation is a task of the sort that EAM and EDP devices are good at, providing someone can write instructions for them.

However, libraries are evolving rapidly beyond the status of passive archives that scholars browse in. In science and technology, they are challenged with an ever-increasing urgency. At one moment someone wants a copy of a document that he has already accurately identified. At the next, someone else needs a list of report or article titles relevant to a problem at hand. Half an hour later, the need may be for a degree of assurance that there's nothing in the files that already answers a question that has come up.

It is probably because of the particular suitability of electronic data processing to file manipulation that not too long ago—maybe ten years—there was common conviction that automation would take care of all this. There seemed to be no reason why you couldn't eventually put scientific or engineering information in a machine's memory, as that information was generated, and pull out whatever you wanted whenever you needed it. We now recognize that we cannot yet organize files and inquiries so as to realize this ideal.

Evan Herbert, Associate Editor of *International Science and Technology*, put it this way in a recent article (*3*): "New ways of manipulating data can give instant access to networks of files, but retrieval of information still hinges on the transfer of meanings."

This brings us to consideration of where we stand today in the evolution toward the fullest possible application of computer technology to libraries. For purposes of discussion, let us divide this evolutionary process into three phases.

The first phase is automation of the files simply to replace manual operations with machine operations in performing library tasks of conventional sorts.

Tremendous strides have already been and are in the process of being taken toward automating conventional library operations. Punched-card, EAM systems have long since been adopted by many, many libraries in this country and abroad for a variety of purposes—control over the ordering, receipts, cataloging, and routing of documents, monitoring of lending and recovering them, production and arranging of card catalogs, and the like. Many of these EAM systems are presently being converted to EDP.

More recently, more sophisticated automation efforts have been undertaken. Each of the three National libraries—agricultural, congressional, and medical—has completed extensive studies of automation possibilities and automated various operations to varying degrees. The Library of Medicine, for example, produces *Index Medicus* from machine-stored, sorted, and printed records. The National Agricultural Library is embarking on a broad program of mechanization. The Library of Congress already turns out its listing of new serial titles from an automated file, and has recently embarked on a pilot study (sponsored by the Council on Library Resources) of the feasibility of making catalog data in machine-readable form available to all libraries.

Actually, this kind of automation is so widely and rapidly being introduced as to outstrip anyone's ability to describe the present state of affairs accurately. The Documentation Division of SLA and the Library Technology Project of ALA are currently engaged in a survey of what normal library functions, such as "payroll accounts, catalog card production, KWIC indexes, serial records, union lists, circulation control, and current awareness services," individual libraries have automated or plan to automate. When the results of this survey are in, we will know better where we stand; but there is no question that the first phase of computer application to libraries has been completed in the sense that the know-how has been developed, even though problems of interchange of machine records remain and may even be exacerbated by the present uncoordinated stampede toward computers.

The second phase of library automation is the one in which we are today. It involves the question of which documents in the file contain information on a specific subject. In the first phase, we have automated our inventory control over packages of information; in the second, we are trying to automate the process of determining which packages in the inventory contain answers to particular questions. Like the bibliographic inventory control just referred to, this process of identifying items by information content is something that librarians have been doing for years. However, the enormous increase in the body of knowledge, the increase in variety of uses for it, and—particularly in science and technology—the demand for speed and flexibility of retrieval have overtaxed the conventional man-based systems. After all, the file in which the relevant items are sought may be a huge one of 10^6 items or more—and note that the word here is "items," not "bits."

But the effort is still one of performing conventional operations on the files, although in the interests of getting the most possible use out of machines many innovations in file organization have been tried. The best known of these, of course, is coordinate indexing.

The snag that has caught us in this phase is the "transfer of meanings" difficulty referred to by Herbert. Viewed from a slightly different angle, it is also called the "natural language" problem. Many automated systems have subject as well as bibliographic tags attached to the information in their files, but subject searching by unaided computers is so far considerably less than perfect. At the moment, it seems more important to identify what kinds of aid computers can offer in finding what kinds of information than to try to turn the whole job over to computers. The emphasis is on machine-aided rather than all-machine systems, and possibly it will stay there for a long, long time.

Libraries and Machines—A Review*

The application of computers to library operations is discussed in broad terms. Three phases in automation of libraries are identified: the mechanization of conventional operations such as bibliographical control processes and administrative monitoring systems; the automation of search processes based on subject matter; and the move toward new and different kinds of services that computer technology may make possible. We are in the second phase, and snagged by the difficulty experienced by computers dealing with natural language and subjective ambiguities. To move forward through phase two will require a better dialog capacity between man and machine than presently exists. Before progressing into the third phase a better identification of the purposes that our files of information are to serve will be needed.

Practical considerations affecting computer adoption by libraries are identified as: the need to stay in business during conversion to automated modes; the necessity for demonstrating in advance the economic advantages of conversion; the difficulty of proving in advance that conversion will meet real user needs; and the standardization and compatibility problems that will have to be solved to make the various automated libraries able to use one another's services efficiently.

BURTON W. ADKINSON AND CHARLES M. STEARNS

Office of Science Information Service
National Science Foundation

This is an attempt to identify in general terms the point that has been reached in marrying computer technology and library operations, and to sketch the more important possibilities and problems that lie ahead as the process continues. Rather than exploring technical questions of specific computer potentialities in specific library applications, it views the problem very broadly, and with emphasis on the library rather than the computer partner in the prospective marriage.

A recent Systems Development Corporation publication (*1*) contains the following statement:

> Modern information technology has made it possible to place much of the accumulated knowledge of the human race within reach of a man's fingertips, so to speak. But the capacity of executives, scientists, and scholars to absorb information has not increased. Therefore, as the amount of available information grows, there is a parallel need for a more precise capability to retrieve specific data in any area of interest.

A considerable percentage of the accumulated knowledge referred to passes through or into libraries. Their problem is determination of file systems and file management to provide this capability. A library is essentially a set of files. For example, the Library of Congress has a basic set of files that in all contain some 42 million items. Thirteen million are books, serials, bound volumes of newspapers, and the like; 18 million are manuscripts. Then there are 21 million more items in files of maps, microprints, music, photographs, and so on (*2*).

Means of access to these collections are files. The basic Library of Congress catalog card file has about seven million titles in it, but this covers only parts of the document collection. Beyond the basic file are the files that permit access by title, subject, author, etc. Beyond these again are collections of documents, such as abstracting and indexing journals, that are the search tools needed to approach the other files; and of course, the files of cards needed to get at the search tools. Then there are what you might call administrative files—records of documents on loan, or ordered, or being bound, and what not.

Of course, any one problem may require reference to many of these files. Furthermore, any change, such as subscribing to a new journal or just receiving the latest issue of an old one, is like a rock thrown into a lake—it sends a ripple of change across all the surrounding interrelated files.

* Based on a talk delivered at the 1966 Spring Joint Computer Conference in Boston, Mass.

To move forward in this phase we need improvement in the man-machine interface. One way of increasing the usefulness of computers in this aspect of information retrieval is to develop a computer configuration that lets the user play a 20-questions sort of game, a real-time dialog that zeroes in on agreement as to just what information in the file most closely corresponds to a need of the moment. If the computer learns in the process so as to enter the next dialog with better insight, so much the better. But the contribution of computers to information retrieval is going to remain disappointingly small until the dialog requirement is met.

In the third phase of evolution toward mechanized information services, conventional operations may disappear almost completely, and the storage and search will be for information itself without regard for the item or document that contains it. As one toiler (*4*) in this vineyard recently put it: "The ultimate goal seems to have been posed—bookless libraries."

Here we need clearer understanding of the purposes that our files of information are to serve. We must grasp, for instance, the desirable difference in structure and manipulation of files to be used here and now as opposed to archive files to be used by posterity. Libraries and computer experts together are going to have to explore the future requirements for filed information and the implications of these requirements for computer applications.

One thing is clear. The ultimate system is going to require the capability to thread through files of greater variety than any now known and from a greater variety of angles to attack than has yet been achieved. Every file, subfile, and superfile must be interconnected so that any and every one of them can be queried at will, with any combination of them brought to bear on a given problem. To risk an example: answering a single question may require fast display of a single page selected on the basis of subject, age, author, availability, restrictions on use, language, accuracy, prior and subsequent or related information on the same subject, or any combination of these. And, the information sought may not even appear anywhere in printed form on any page.

This phase we have not really entered. Some experimental steps are being taken toward it, however. Project MAC at MIT, which involves shared time use from remote consoles of a pool of information that is heterogeneous from more angles than can easily be listed, is one example (*5*).

The point we would like to make here, though, is that either by foresight or experiment the purposes that future files and retrieval systems must serve, and the roles that computers can play, will have to be worked out by the library and computer communities in close conjunction if we are to move through phase three at all.

Turn now to certain practical considerations that are also going to affect both the degree and the rate of computer adoption by the library community, in all three phases.

First, libraries as they evolve must stay in business as they do so. They cannot shut down for retooling for a next-generation computerized model.

Second, mechanization of old capabilities will not likely occur unless it guarantees large economic gain over conventional ways of doing things.

Third, new capabilities will not meet with an eager reception until someone offers compelling proof that they will meet a real user need.

Fourth, the largest pay-off from automation will only be achieved when many kinds of computer-readable records can be freely interchanged among individual libraries, and this introduces the ancient stumbling-block of standardization.

It is important not to underestimate the basically conservative attitude that underlies these four points. It arises out of a long history of financial starvation of library management. It has been intensified in recent years by the exponentially increasing amount of information that has to be obtained and filed, and the increasing library manpower and space that this entails. Any additional subsidizing of libraries that results from recent legislation may help to erode library conservatism. However, the tendency will be to change cautiously, and to use whatever money becomes available for the conventional purposes that have been so hard to achieve in the past with limited cash—buying more books and building more space.

With regard to the first of the four points, not much needs to be said. Regardless of what use any one individually may make of libraries, they really cannot go out of business for any appreciable length of time. No matter how logical it may be to put the whole card catalog on magnetic tapes, nobody is going to be willing to get along without the cards while the tapes are in the making.

The second point was the need for economic justification of change in established ways of operating. This is not always as simple as it may seem. There is psychological difficulty in scrapping an enormous investment in conventional tools and training. There is logical difficulty in demonstrating a favorable cost-benefit ratio when the benefit rests on some still unknown quantity. No simple assertion that computers will do something cheaper is likely to persuade librarians. They have a right to be suspicious, and they are.

Thirdly, when it comes to proposing computer applications to achieve capabilities beyond the present ones, there is another barrier. One of the commonest allegations in the science information business is that scientists and engineers only accept conventional services because they haven't had access to better ones. But, in spite of years of effort and hundreds of thousands of dollars spent on trying to identify the *real* needs of users of scientific information, we remain unable to describe them to anyone's satisfaction. We are unable even to predict whether a new service would find customers even if it *could* be demonstrated that it satisfied a need.

Finally, with respect to the need for standardization, it is clear that part of the necessary economic justification referred to above will rest on increased capability to engage in cooperative load-sharing arrangements. However, besides challenging the historically derived bias toward local self-sufficiency, this will require compatibility in a broad area ranging from cataloging rules to machine formatting.

It follows that proposals to set up computer-based operations that offer new and appealing services may not meet as kind a reception as expected. Furthermore, the same argument works in reverse. Proposals that result in dropping or curtailing past services find hard going. A case in point is the opposition that computer-stored coordinate indexing, or reduction of files to microstorage, meets when someone realizes that the time-honored pastime of "browsing" will be threatened.

These obstacles are not insuperable; but they need to be recognized more clearly. Evolution toward better libraries with different roles will take place as computer applications are made, as the progress through phase one and into phase two already has shown. To speed the process through phase three simply requires that the computer community approach realistically the difficulties that face the library community and join with it in spelling out both the services that the systems of the future must provide and ways of realizing them.

Note: Time limitations on the talk on which this paper is based precluded extended discussion of present operations and experiments in the introduction of computers into libraries. Actually, there is such an enormous amount of activity in this field that one seeking representative examples is confronted by an embarrassment of riches. This profusion of examples may be discovered in such bibliographies as the one compiled by Edward M. McCormick for the University of Illinois Clinic on Library Applications of Data Processing, (*6*) or the several contained in the proceedings of the conference on libraries and automation held at Airlie Foundation (*7*).

References

1. The National Information Problem, *SDC Magazine,* 9: 1–15, (February 1966).
2. Library of Congress, *Automation and the Library of Congress,* Washington, D. C., 1963, p. 7.
3. *International Science and Technology,* (March 1966), p. 26.
4. Learning How to Live Under Water, *The Bookmark,* New York State Library, 25: 37–49 (November 1965).
5. The MAC System, *IEEE Spectrum,* 2: 56–64 (January 1965).
6. University of Illinois, Graduate School of Library Science, *Proceedings of the 1963 Clinic on Library Applications of Data Processing,* 1963.
7. Library of Congress, *Libraries and Automation,* Washington, D. C., 1964.

A Documentation Training Model*

This contribution reports on the design and development of a series of model information retrieval and library systems. These are designed to allow documentation students access to a variety of basic files, permit lecture demonstrations, enable comparisons (since all files relate to the same collection), and to serve as the depository for the input assignments of advanced students.

Comments are provided as to the choice of systems made, and their potential applications for research.

C. D. BATTY, B.A., F.L.A.

Head of Department of Information Retrieval Studies, College of Librarianship, Wales.

The training model to be discussed is based on an integrated set of manual and mechanised indexing systems, all handling the same body of information from a limited subject field. By extending the scope of the model's operations to include prior and subsequent activities like the selection and abstracting of the documents to be indexed, and the preparation and dissemination of material through the use of the indexes, the model may be used for a wide range of documentation training, principally at three levels: demonstration by the lecturer to the students; use by the students in the retrieval and dissemination of information; and development by the students through the selection and abstracting of documents, the indexing and storage of information, and ultimately the use of feedback from the dissemination stage to improve the systems.

There are principally two reasons for the development of such a model. The first concerns access to systems in working libraries; the second concerns the usefulness of working systems to professional education. Professional training has two closely linked aspects; theoretical training belongs in the classroom—practical training in the working library. But while this division is fairly satisfactory in areas such as reference work, administration and special areas of attention like children's libraries and music libraries, training in documentation and indexing presents a different situation.

Few document indexing systems in special libraries are readily available in the vicinity of most library schools and even though some may be, it is most unlikely to find a complete range of all types of system. More conclusively, libraries of this type are highly conscious of efficient performance; they will therefore be the more reluctant to allow students ready access to their systems, and will hardly ever allow whole classes of students to operate and experiment with them. In any case the instructional value of existing documentation systems in libraries is lessened by the wide range of subjects covered collectively by all the systems encountered. The inevitable unfamiliarity of many of the subjects will limit reliable assessment of a system's efficiency and the variety of subjects will effectively inhibit that comparative view which should provide the best instruction of all.

It is possible to go further and say that in general a complete working system is unnecessary for instructional purposes. There is a minimum size, it is true—but once this has been achieved, further development is a dispensable luxury except where the operations involved constitute the instruction in hand. Rather than use existing documentation systems, then, it may actually be better to construct a set of representative systems within the library school. This would ensure: first, a full range of possible systems; second, constant, immediate, and unrestricted access; and third, better facilities for comparative study, since all systems could handle the same material. Certain economies of time and effort would also be possible where systems shared operations or equipment, for instance, in the selection or abstracting of documents, in devising languages for similar indexing systems, or in the use of equipment handling information recorded in a similar physical form.

The reason why such a model might seem in any way novel is partly historical and partly financial. In the United Kingdom a concentration in the past on public

* The basis of this article was a paper read at the FID Congress, Washington, 1965.

library needs and standards, which even now affects the teaching of more specialized disciplines and techniques, and the inclusion of library schools in colleges of commerce and technology (which belong to municipal authorities) rather than in universities has meant not only that expensive equipment is low on any list of priorities but, more important, that little staff time is available for the development of new training methods.

Only recently have a few library schools broken away from this pattern, notably the post-graduate school in Sheffield University, the library schools in universities in Scotland and Northern Ireland, and the College of Librarianship in Wales, an independent college concerned only with library science.

The model under discussion has grown in concept and execution from a practice common enough in British library schools: that of constructing miniature catalogues from the entries composed in practical cataloguing sessions. These, without much trouble, can be made to form parallel dictionary and classified catalogues, not only to give the students practice in their construction and manipulation, but also to offer a comparison of types. An easy translation into the field of documentation and indexing was made when the new Library Association examination syllabus offered such papers as the Handling and Dissemination of Information, Special Librarianship, the Theory of Classification, and (in the post-graduate syllabus) Indexing, Abstracting, and Information Retrieval. At the same time the new syllabus encouraged attention in this field in broader, though still relevant papers, like The Organisation of Knowledge.

In a paper to the 1964 annual conference of ASLIB the author described work involved in the construction of a simple machine sorted punched card index to the periodical literature of a limited field and considered some problems of its demonstration and use in the teaching of mechanized information retrieval. The establishment of the College of Librarianship in Wales offered occasion for the rationalization of work and ideas arising from this project. Temporarily within the limits of the Library Association examination syllabus, but encouraged to look ahead to more advanced courses, discussion began on the provision of a whole range of depth indexing methods, from a conventional classified card index to computer systems, and on the possibility of extending this effort into the wider field of documentation by selecting and preparing material, and disseminating it, all as part of the same sequence of operations.

The material used as a basis for the model is the literature on information retrieval, because of the familiarity of the concepts and terminology, the limited size of the field, and therefore of the index languages, and the existence of *Library Science Abstracts* as a predigested form from which the students may work and to which easy reference may be made. And since *Library Science Abstracts* did not begin until 1950 there remains a considerable body of material requiring selection and abstracting.

The model is being built up in three stages.

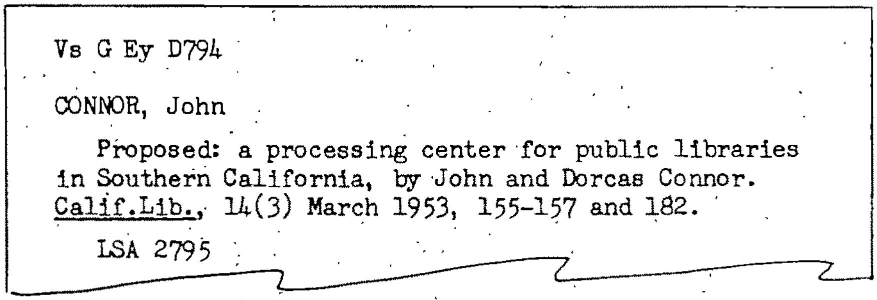
The first stage comprises manual systems. Two of these, a conventional classified catalogue and a rotated classified file in visible index form, use a new scheme of classification for library science devised by the Classification Research Group in London.

The same scheme is used for the College library and is its first practical application (beginning actually before its publication) beginning in March 1966. After some 15 months of experience, the College began to circulate CRG/EXTRA (EXTensions, Revisions, and Alterations) to outline problems encountered and solutions proposed. It is hoped in this way to secure comment and discussion on use to assist the Classification Research Group in the work on a second edition. It is a faceted scheme, covering all types of library and library service, and the materials and techniques of librarianship, as well as general questions affecting the profession and a wide variety of fringe topics, from education to the book trade. All facets (or classes) are listed in the schedule in general to specific order, and combine retroactively to give the citation order for elements in compound class numbers: libraries—services—materials—processes—general questions. The classifier must be able not only to analyze abstracts into component terms (as for coordinate indexing) but also to recognize an order of significance. Once this has been done the application of the scheme is simple.

The conventional catalogue is on 125×75mm cards and the entries show the class number by which they are filed, author, title and source of the article, as well as the abstract number. There is an alphabetical subject index to the main classified file constructed according to chain procedure, where each element (each link in the hierarchical chain) is indexed with its appropriate superordinate terms used as descriptors, to provide relevant access for inquiries initiated at too general a level. See Fig. 1. About five hundred abstracts have so far been classified with an average of about five "elements" in each class number.

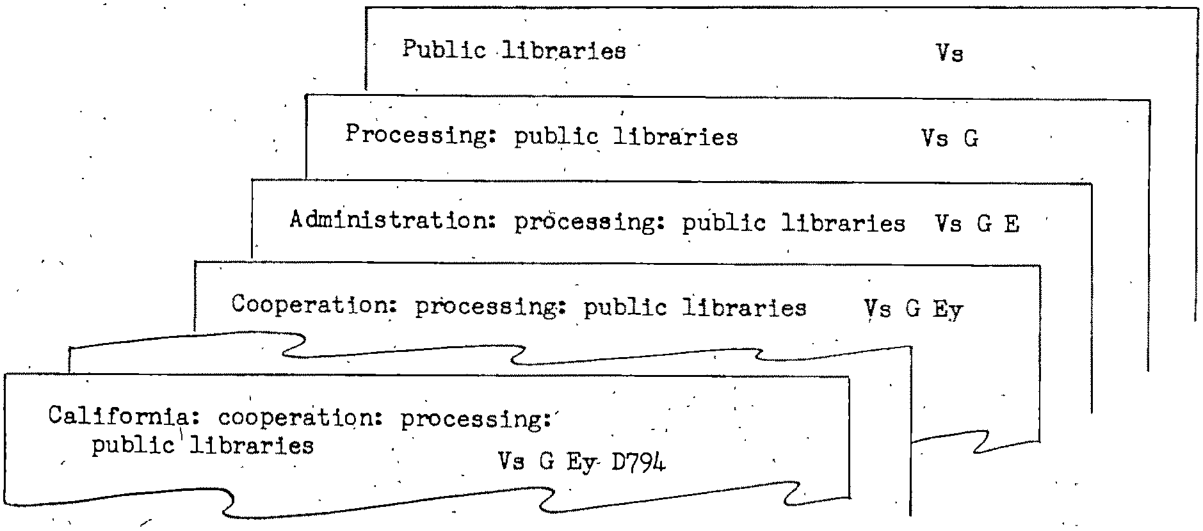
The nature of the classification scheme has prompted the inclusion in the model of a variation of the rotated classified file. Instead of *rotating* the elements in the class number, however, (so that ABCD becomes in turn BCDA, CDAB, and DABC on four different entries) this index adopts what actually happens in many so-called rotated indexes produced on computers, where the entire line is *shifted along* to the left, one element at a time. This provides multiple entry as before, since each element in turn appears at a marked filing position, but it does not disturb their order, thus preserving the classified heading as a coded analysis of the whole subject of the document and an address in the classified shelf order.

This index uses Kalamazoo "Copystrip" binders, each containing twenty-five dividers with thirty 2-line entries per page, and each entry showing the class number, author, and brief title, and the abstract number. See Fig. 2.



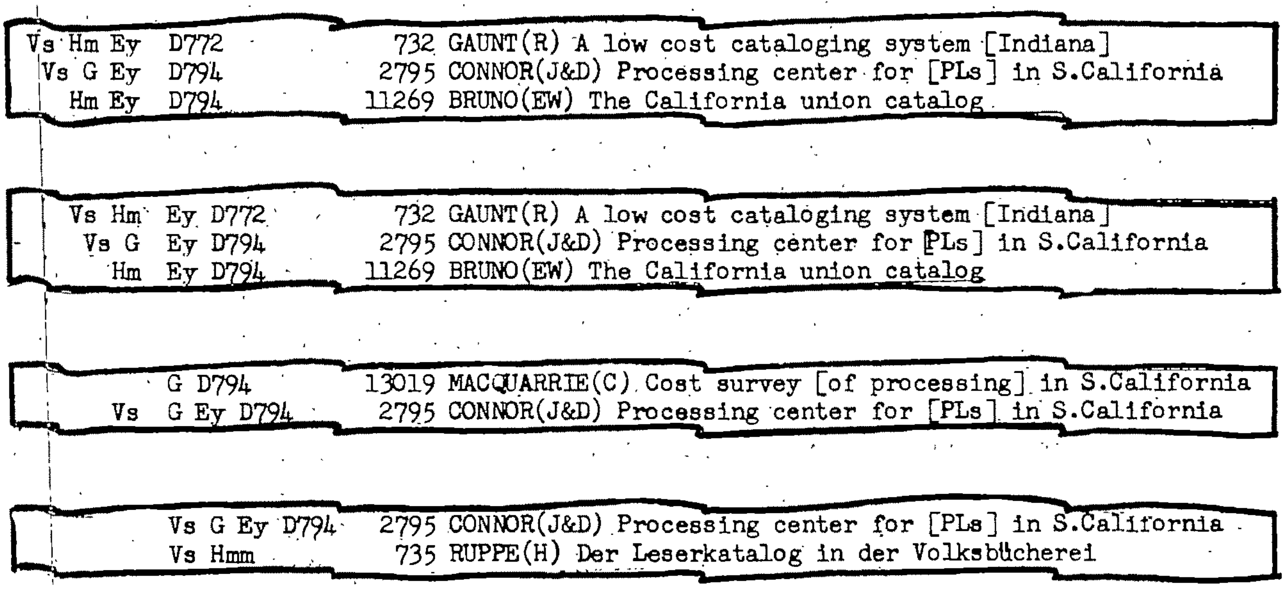
```
Vs G Ey D794

CONNOR, John

   Proposed: a processing center for public libraries
in Southern California, by John and Dorcas Connor.
Calif.Lib., 14(3) March 1953, 155-157 and 182.

   LSA 2795
```

FIG. 1*a*. Main entry for the conventional classified catalogue



```
Public libraries                                  Vs
Processing: public libraries                      Vs G
Administration: processing: public libraries      Vs G E
Cooperation: processing: public libraries         Vs G Ey
California: cooperation: processing:
   public libraries
                                   Vs G Ey D794
```

FIG. 1*b*. Alphabetical subject index entries by chain procedure for the main entry shown in Fig. 1*a*



```
Vs Hm Ey  D772          732 GAUNT(R) A low cost cataloging system [Indiana]
 Vs G Ey  D794         2795 CONNOR(J&D) Processing center for [PLs] in S.California
    Hm Ey  D794       11269 BRUNO(EW) The California union catalog

   Vs Hm  Ey D772       732 GAUNT(R) A low cost cataloging system [Indiana]
     Vs G  Ey D794     2795 CONNOR(J&D) Processing center for [PLs] in S.California
        Hm  Ey D794   11269 BRUNO(EW) The California union catalog

          G D794      13019 MACQUARRIE(C) Cost survey [of processing] in S.California
    Vs    G Ey D794    2795 CONNOR(J&D) Processing center for [PLs] in S.California

          Vs G Ey D794 2795 CONNOR(J&D) Processing center for [PLs] in S.California
          Vs Hmm        735 RUPPE(H) Der Leserkatalog in der Volksbücherei
```

FIG. 2. CELT index (Classified Entries in Lateral Transposition)

The other manual systems are all coordinate indexes, based on the same index language, though not using it in identical ways. Two of them are term entry systems: respectively number matching and optical coincidence. At first it was intended not to include in the model more than one system of any type (in this case manual systems based on term entry) but to construct small static models simply to demonstrate the different devices through which the same basic system might be manifested. An optical coincidence system was obviously better for the model because of its ease of operation, with a number matching system as its static counterpart, but since in practice it proved easier to construct the number matching system first and transfer it *en bloc* to optical coincidence feature cards it was decided to keep them both. The number matching system uses standard 200×125 mm index cards filed in alphabetical order of the index-words; the optical coincidence system uses Manisort equipment, with cards of 10,000 document capacity. It has been observed that abstracts require an average of eight terms for adequate description.

The third manual coordinate index is an item entry system using Cope-Chat standard P.1. 75-hole cards, 6×4 in. For this system the thesaurus is given a random number coding of pairs of two digit numbers; since the whole card is regarded as a single field the total notation available is $^{75}/_{2}=2862.5$ places—ample for coding the estimated 500- or 600-word thesaurus. The cards show only the abstract number and the serial numbers of the index words, together with the random number coding. It is physically possible to include the whole abstract on a card of this kind, but in the model it was felt that to duplicate the master file of abstracts and to make one index independent would be to the disadvantage of the model as a whole. See Fig. 3.

Stage two in the development of the model concerns mechanized systems using 80-column punched cards and ICT equipment, comprising an alphanumeric keypunch, a verifier, an ICT 302 sorter, an interpreter, a mark sensing and reproducing punch and a tabulator-printer. Most of the work on the punched card indexes has been done on the mark sensing punch (so that students could prepare the cards themselves), and the sorter. The tabulator-printer lists abstract numbers of cards retrieved in response to a particular search. All these systems are item entry systems.

The first is the simplest of all, using positional punching for serially numbered index terms. These are punched into a single field of seventy columns. A second field comprising the last 10 columns is reserved for the abstract number; at present this is the *Library Science Abstract* number of up to five digits, punched into the last five columns only, but if, as is intended, an independent file of abstracts of pre-1950 material is built up, then distinguishing prefixes may be used in the earlier columns of that field. See Fig. 4.

The second punched card system is simply a mechanized form of the edge notched card index mentioned above, with the difference that the pairs of two digit numbers used in the random number coding are divided between two fields of 100 positions each (the first two digit number of each pair in the first field and the second number in the second field). This increases the notational capacity to 10,000 codings and, by a wider scattering of index terms, lowers the proportion of false drops. Cols. 1–10 and 11–20 are used for the subject coding; cols. 21–70 are reserved for author/title and/or source, though this field has not yet been used. Cols. 71–80 are reserved for the abstract number as described above. See Fig. 5.

The third is experimental, involving categorization of the index language, and a consequent categorization of the coding for the punched card. For obvious reasons this is not yet a part of the model and may never be. It is a useful adjunct that feeds the model with ideas. One version designed for 40-column cards and equipment categorizes the index language into three main fields of library, material, process, and a fourth field of the names of people and places. A mixture of direct and condensed coding and the combination of zone and digital punchings has substantially increased the capacity of the smaller card.

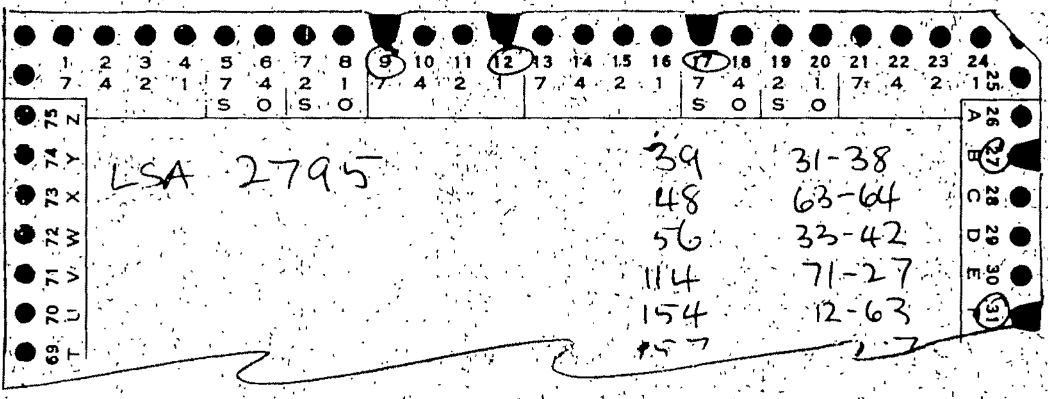


FIG. 3. Edge-punched card using random coding of pairs of two digit numbers

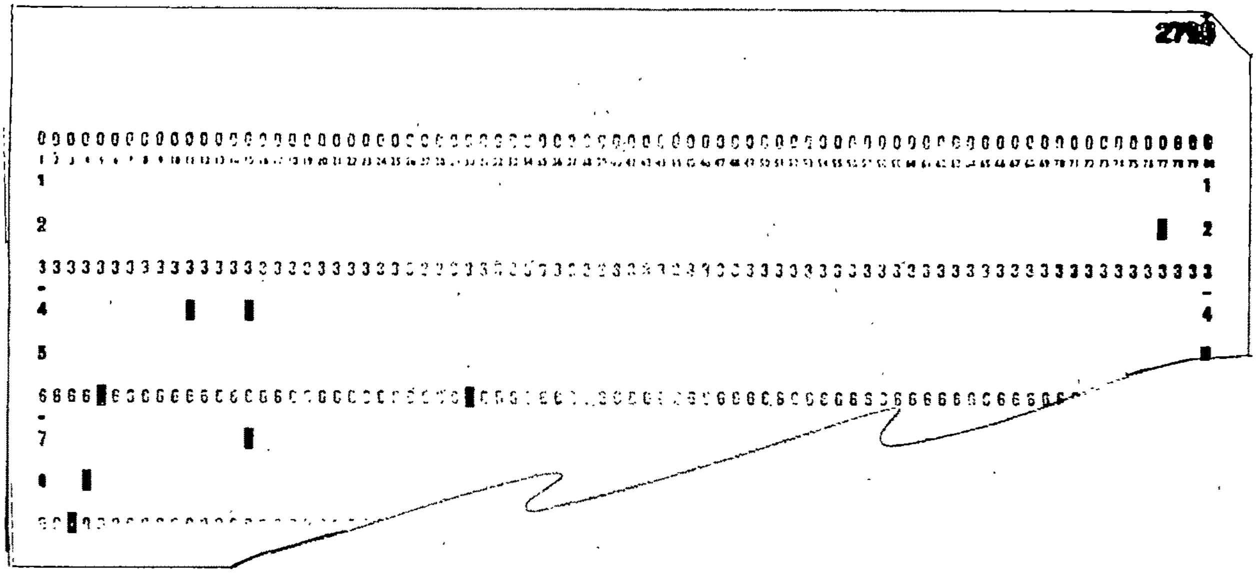


FIG. 4. Item entry system on 80-column cards using direct coding and positional punching

It might be relevant to say something here about the coordinate index thesaurus. This is basically the same for all systems. It was worked out originally for the first of the mechanized systems just mentioned—the one using serial positional punching—by the empirical method of indexing as many abstracts as possible and collating the terms used. An early assumption that a thesaurus for information retrieval and documentation would barely exceed 500 terms (and that therefore positional punching could be considered) has been borne out in practice, though there are certainly more terms to be added. The thesaurus is maintained on punched cards for control and easy production.

Stage three concerns additional or substitute systems considered or planned for a computer now under consideration for the College, to be used on a time-sharing basis with other institutions in the area. These computer systems include a classified file based on the conventional manual classified file, to be developed as a classified catalogue programme, a serial file (item entry) for sequential scanning, and an inverted file (term entry) searched through a random access device. In addition to these, which develop naturally out of systems in the first and second stages of development, it is intended to add a KWIC index and a lattice index. The latter is a type comparatively little studied as yet and should provide an interesting field for research.

In addition to the indexes already discussed there is a master file of 200–125 mm index cards containing all class numbers, index entries, and codings. This is filed in abstract number order. See Fig. 6.

The model was "primed" with 1,000 documents coded in each system. This work was done by the staff in order to test and develop the thesaurus and the classification scheme, to establish forms and routines for student work, and to build the model to a size sufficient for initial demonstration by the lecturer and manipulation by the student. From this point work by the

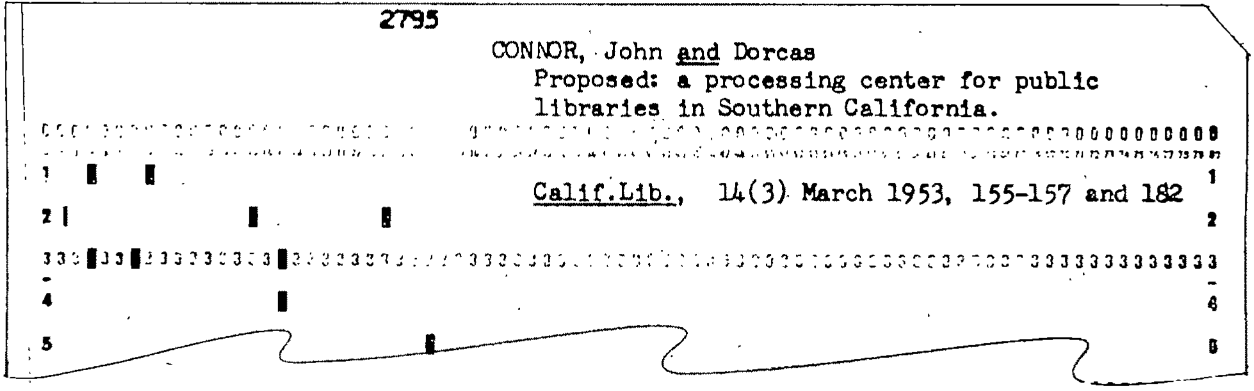
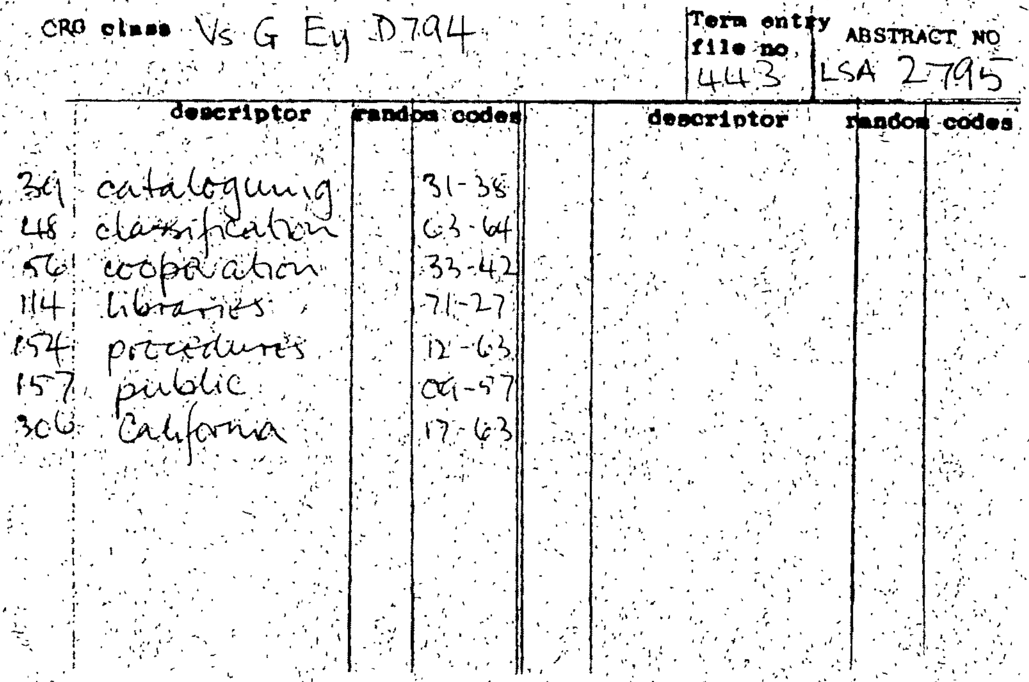

2795
CONNOR, John and Dorcas
Proposed: a processing center for public
libraries in Southern California.
Calif.Lib., 14(3) March 1953, 155-157 and 182

FIG. 5. Item entry system on 80-column cards using random coding in two fields



CRG class Vs G Ey D794

Term entry file no. 443

ABSTRACT NO LSA 2795

	descriptor	random codes		descriptor	random codes
39	cataloguing	31-38			
48	classification	63-64			
56	cooperation	33-42			
114	libraries	71-27			
154	procedures	12-63			
157	public	09-57			
306	California	17-63			

FIG. 6. Master card showing all codings, index terms, and class number; filed in abstract number order.

students began to develop and extend the model, controlled and occasionally supplemented by the staff. Staff participation is now concerned to keep all parts of the model at the same stage of development.

It has been the author's present concern to describe the training model itself; it is hoped at not too distant a date to offer some account of its use and value. It might be useful at this stage, however, to indicate in general terms the uses envisaged for the model and which it already serves.

Elementary classes in the organization of knowledge receive demonstration by the lecturer of the indexing systems and the pattern of documentation activity that includes them. At this level little practical work is possible or even desirable, since the classes are large, the courses fairly general, and the students unprepared.

Second year classes on relevant courses use the model as an object lesson on which to base seminar discussion and to gain experience in its manipulation, setting up programmes for demand searchers, and the general production of bibliographies.

Classes specializing in documentation, such as in the handling and dissemination of information, proceed from manipulation to development of the model. At this stage students begin, by indexing or classifying abstracts for the model, coding them for each system in turn, and go on to the preparation of abstracts for the indexer.

It is hoped at a later stage to use the model for theoretical training in SDI systems, but this, like so much that may be expected from such a model, is still part of the future.

An Experimental System for Automatic Identification of Personal Names and Personal Titles in Newspaper Texts

Natural language seems to contain various special-purpose subsystems, e.g., personal titles, personal names, dates, street addresses, place names—each with its own structure which relative to the total structure of language is rather simple.

An ability to identify automatically words and word strings belonging to various special-purpose linguistic subsystems (akin to some thesaurus classes) may prove to be very useful since they play an important role in the making of indexes and in various systems for extracting and distributing information.

This article describes some of the main problems involved in automatic identification in newspaper texts of words and word strings belonging to two important linguistic subsystems, *viz.*, personal titles and names; lists some of the major rules of an algorithm designed to perform this task; presents statistics concerning the algorithm's accuracy and exhaustiveness obtained in manual application of the algorithm to texts; and suggests some applications for computer programs capable of recognizing personal titles and names.

The results obtained indicate that an automatic system capable of accurate and exhaustive identification of personal titles and names in texts requires recognition procedures which are rather complex.

It is therefore suggested that along with researching and developing methods for high-quality *automatic* classification of words in texts, it may be advisable to set up efficient procedures for *manual* classification and tagging of words in texts, and *automatic* extraction of data from texts which were recognized either manually or automatically.

Such action seems appropriate since *automatic* extraction of information from *manually recognized* texts would probably constitute a valuable service, and, when *automatic* procedures for identifying dates, personal names, personal titles, trade names, company names, chemical formulas, numbers and measure words, and so forth become competitive with manual ones, the data-processing profession will be already in possession of operational computer programs capable of extracting data from recognized texts.

CASIMIR BORKOWSKI

Thomas J. Watson IBM Research Center
Yorktown Heights, N. Y.

• Motivation for the Experiment

One of the major questions of the day is the extent to which a computer can be instructed to identify various parts of texts written in plain, ordinary language. In trying to answer this question, we set ourselves the preliminary limited objective of developing an automatic procedure for identifying proper names in English-language texts and classifying them according to type, for example, as names of persons, names of organizations, names of places, etc.

The selection of this objective was based on the following considerations:

1. Proper names are easier to identify and classify automatically than other parts of texts (a) because of orthographic and style rules (capitalization, etc.) and (b) because relative to the rest of the language their structure is usually quite simple.
2. Identification of proper names can be carried out largely independently of the identification of other parts of texts.
3. Attempts at automatic identifications may increase our knowledge of the structure of language.
4. An ability to identify proper names automatically may prove to be very useful since proper names play an important role in the making of various indexes as well as in various other systems for extracting and distributing information. Automation of name identification might therefore present a certain economic advantage.

Since automatic identification of names of persons appeared easier than that of proper names of other types, we decided to commence our investigation of proper names with an investigation of personal names.

In newspaper texts, personal names are often preceded by personal titles, that is, words or phrases bestowed on individuals as a mark of distinction, rank, or dignity and frequently describing or implying office or vocation.

Because personal titles provide important information about the persons bearing these titles and furthermore because automatic identification of prenomial titles helps in automatic identification of personal names, we combined both automatic identification procedures into a single system.

A newspaper was selected as the source of texts because many newspapers are printed by means of typesetting tapes which could be converted to computer-legible form; because automating the extraction of information from newspaper morgues (files of reference material) presents a challenging problem; and because newspapers contain a great variety of personal names and titles.

• Some Problems of Automatic Name Identification

Automatic identification of names of persons in texts is of course not without its difficulties. First of all, many personal names are homographic, that is orthographically identical, with other types of words in the language. This is the case since among the main sources of surnames are titles, e.g., *Baron, King;* names of occupations, e.g., *Baker, Smith;* topographic terms, e.g., *Bridges, Dale;* names of animals, e.g., *Bull, Fox;* names of places, e.g., *Danzig, London;* names of plants, trees, etc., e.g., *Bean, Bush;* personal attributes, e.g., *Stern, Wise,* and so forth.

There is considerable homography between personal names and place names due to the fact that not only are the names of places a frequent source of personal names, but because many localities were named after people, as, for example, Berkeley, California, and St. Augustine, Florida. And to make matters worse, hotels and business firms can be named after people and are often referred to by an abbreviated name which is that of a person, e.g., "I am staying at the Mark Hopkins," "Ford was hit by a strike last week." As for personal names like *Elizabeth Arden* and *Max Factor,* they designate persons as well as business firms, while *Philip Morris* is the name of a person, of a corporation, and a brand of cigarettes.

Yet another difficulty arises in the case of names of persons, e.g., *Madison, Sir Francis Drake,* when they perform a naming function with regard to something, say an avenue, e.g., *Madison Avenue,* or a hotel, e.g., *Hotel Sir Francis Drake.* Presumably, it would be worthwhile to distinguish automatically references to persons from references to things named after persons.

Still another difficulty in recognizing personal names results from the fact that personal titles are not unfailing aids in identifying personal names because titles themselves can be homographic with other types of words. For instance, *General* is a military rank in *General Mobutu* but not in *General Motors.*

Further difficulties result from the fact that some titles are homographic with given names. It is not a simple matter to specify the rules that would enable an automaton to decide when *Dean* is a personal name and when it is a personal title, e.g., *Dean Rusk, Dean Wiesner.*

• State of the Art

The name of the science dealing with the origins, forms, and usage of proper names is onomastics. Although onomastics has important contributions to make to many problems connected with computer identification of names in texts, at present there seems to be little interest among name specialists in computation problems such as automatic recognition of names.

An interesting and—as far as we know—the only previous experiment in computer recognition of names which has been described in the open literature was performed by a documentation specialist, Professor Susan Artandi at Rutgers School of Library Service (1) (2). Professor Artandi used two relatively simple methods of identifying proper names in a pioneering experiment whose goal was to determine the extent to which a computer could assist human editors in indexing documents with whose subjects the editors have only a minimum amount of familiarity.

The first method extracts from texts capitalized words and strings of capitalized words while the second method records "all capitalized words which appear in the document text with the four words preceding it, the four words following it, and its page number." The proportion of "useful indexing entries" which was produced by these methods was about 50%.

• Experimental Design

Our procedure in setting up an automatic method for identifying personal names was approximately as follows:

1. We investigated permissible patterns of personal titles and of English, Spanish, Russian, Chinese, and other personal names whose occurrence in texts we could anticipate.
2. We searched the literature for information concerning the structures of personal names and titles, methods of processing texts, and other data pertinent to automatic identification of personal names and titles in texts.
3. We obtained a 60,000-word sample of newspaper texts and determined patterns of occurrence of personal names and titles in texts; patterns of personal names and titles occurring in texts; and problems involved

in distinguishing personal names and titles from each other and from other parts of texts.

4. Based on investigations 1, 2, and 3, we set up an automatic procedure designed to identify personal names and titles in newspaper texts. This procedure was embodied in flow charts and a dictionary.

5. We tested our procedure manually on a 100,000-word sample of new newspaper texts, and we amended the rules and expanded the dictionary on the basis of the information provided by the tests.

6. We then stabilized the improved procedure, tested it out manually on a new 40,000-word sample of new newspaper texts, and collected statistics concerning its accuracy and exhaustiveness. (Our reasons for applying the algorithm manually were as follows: Our identification system was embodied in dictionary entries and flow charts which were sufficiently detailed to permit accurate execution of recognition procedures, and we thought that it would not pay to code and debug over a period of months what would probably turn out to be a "one-shot" program.)

7. We then investigated what types of errors had occurred and proposed various amendments to the automatic recognition procedure.

8. We speculated about possible applications of a computer program capable of recognizing names and titles of persons in newspaper texts.

We selected recognition rules, or rather recognition hypotheses, which are basically quite simple with the intention of finding out how many correct identifications and how many errors they produce.

We plan to amend these hypotheses in the light of the results obtained. Refinements, additions, reformulations of the rules, as well as changes in basic methodology, will be brought in as required and the trade-off between the complexity and the effectiveness of the rules will be noted.

In other words, the present set of identification rules is a first approximation, a first scratch of the surface. We think, however, that the discovery of the area in which the rules fail will be helpful in suggesting new directions and methodology for research on automatic identification of parts of texts written in ordinary language.

The basic assumptions made by the rules which we are about to describe may seem quite unsophisticated. They concern the meanings of words, of phrases, of affixes, of punctuation marks, of capitalizations, etc., which are encountered in newspaper texts.

• The Goals

The goals of our initial system for automatic identification were quite modest. The solution of many difficult or complex identification problems was not attempted. For instance, we did not attempt to differentiate the names of persons from the names of various objects named after persons, e.g., *a Garand;* the names of fictional and symbolic characters, e.g., *Simon Legree;* the names of horses and other animals, e.g., *Uncle Max, Vicar Hanover,* and so forth.

Likewise, we did not attempt to specify the rules for separating into different strings adjacent names, e.g., *Winifred Beethoven,* which will occasionally occur in sentences with double-object verbs, e.g., *gave* in "John gave Winifred Beethoven's *Rasoumovsky* numbers one and two"; and in sentences without a comma after an adjunct phrase, e.g., "After his encounter with Thomas Hood had to retreat," and so forth. Such an attempt will be made later if a significantly high number of contiguous names is found to occur in texts.

The solution of some problems of automatic name and title identification may require stronger theoretical assumptions about sentence and text structure and more elaborate techniques of sentence and text analysis. For instance, since automatic parsing of sentences may be helpful in identifying sequences of names each of which is followed by its title, e.g., "The President nominated John Gordon Ambassador to Guatemala, William T. M. Beale, Jr., Ambassador to Jamaica . . .," future automatic systems for identifying personal titles and names may parse sentences containing double-object verbs and strings consisting of personal names followed by titles. However, since parsing and other types of analyses may be expensive, it would seem advisable to apply them only when they can reasonably be expected to provide economic solutions to valid problems.

• Some Identification Rules

Our rules describe the arrangement in the sentence of the words, phrases, and punctuation marks which are pertinent to the identification of names and prenomial titles. Generally, the description starts with the first, that is, the leftmost pertinent element of a sentence and terminates with the last, or rightmost pertinent element.

For greater ease of reading, the rules are expressed here in narrative form. For the sake of brevity, only some of their more important features are listed here. A more complete description of identification rules is available elsewhere (*3*).

Our rules for recognizing names of persons take advantage of the style rules of *The New York Times*. We would conjecture that whereas details of name recognition rules may vary from newspaper to newspaper, their general pattern will remain fairly stable and independent of editorial conventions.

The rule for identifying personal titles which was selected as a reasonable first approximation states that a word or phrase in text is a personal title either:

1. If it matches a word or string of words on a list of titles

or

2. If it matches a word or a string of words which is on a list of words and phrases which commonly combine with titles, e.g., *Acting, Assistant, Vice,* and is followed by a personal title, e.g., *Acting Mayor, Acting Assistant Vice President*

or

3. If it is a personal title followed by a word or a string of words which is on a list of words which com-

monly combine with titles, e.g., *-elect, at Large, pro tempore*, as in *Senator-elect, Ambassador at Large, President pro tempore*.

The rule for identifying names which follow personal titles is likewise quite simple. It states that the capitalized word or string of words and initials which frequently follow a title is the name of a person, e.g., *President Nkrumah, Mr. Green*.

If a title is followed by a word beginning with a lower-case letter or by certain punctuation marks, this indicates that the title is not followed by a name—unless, of course, the lower-case word which follows the title is a name prefix, e.g., *de, von*. The rule states further that the end of a name string which follows a personal title is marked by a word beginning with a lower-case letter or by a punctuation mark (comma, sentence period, dash, semicolon, colon, etc.).

However, a lower-case letter does not mark the end of a name string if the word which begins with it is either

1. A name prefix, e.g., *de* as in *Attorney General Nicolas deB. Katzenbach*

or

2. The last element of a hyphenated Chinese name, e.g., *lai* in *Mr. Chou En-lai*

or

3. The one-letter Spanish word *y*, e.g., *President Jose Bustamante y Rivero*

or

4. If it is one of the Arabic words *ibn, el, al*, and so forth, as in *Abdul-Assiz ibn-Saud, Abd-el Kader, Abd-al-Kadir*.

Important exceptions to this rule are capitalized words and strings of capitalized words which frequently follow personal titles and whose designata are frequently named after personal titles, e.g., *Regent Street, President Hotel, Ambassador Bar, Admiral Transport Company*. Preliminary rules for identifying common namesake words and phrases have been formulated and flowcharted. However, the recognition of namesakes in texts has been given little attention and the rules for identifying them remain very tentative.

While counter-examples to the above rules and to similar rules are easy to invent, the inventors of counter-examples usually miss the point that rules such as these are statistical observations and that in actual application to texts they hold up rather well. Of course, occasionally titles in texts are followed by capitalized words which are not names and for the recognition of which the rules make no provisions, e.g., *British* in "The Prime Minister, British sources said, will arrive on Monday"; and *New York* in "Mr. Stevenson preferred Washington and Mr. Rusk, the Secretary of State, New York." Constructions such as these are, however, quite rare. As stated earlier, among the goals of an investigation of this type is to obtain evidence as to how well the rules hold up and what amendments are required to render them more accurate and to expand their scope.

The rule for identifying the names of persons which do not follow titles states that a string of characters in a text is a name of a person either:

1. If it begins with a capital letter and ends in *-escu, -ev, -icz, -itch, -off, -ovic, -wski, -wska*, and so forth (among the many exceptions to this rule are the words *Itch, Kharkov, Off*, etc.)

or

2. If it begins with the capital letter *O*, followed by an apostrophe, followed in turn by a word beginning with a capital letter, e.g., *O'Brian*

or

3, If it begins with *De, de, Mac, Mc, Von, von*, and so forth followed either directly or after a space by a word beginning with a capital letter, e.g., *De Forest*. (This rule does not apply if *de, di*, etc., are preceded by geographical terms such as *Avenue, Punta, Place, Rio, Rua, Rue*, etc. Among the many exceptions to this rule are the names of localities *Fond du Lac, Juiz de Fora, Santiago de Cuba*, and so forth)

or

4. If it begins with one of the hundred or so common Chinese surnames, e.g., *Chang, Mao, Wang*, and is followed by a string of characters of one of the following patterns:

$¢. . . *P*

$¢. . . #$¢. . *P*

$¢. . . -$¢. . *P*

$¢. . . -¢. . *P*

where $ stands for any capital letter, ¢ for any lower-case letter, where ". . ." indicates that the preceding character may be repeated, where # stands for a space between words, and *P* either for a space or for any punctuation mark other than a hyphen, and where "-" is a hyphen, e.g., *Chu Teh, Chou En-lai*

or

5. If it matches any string of characters which is on a list of names of persons. The names list contains no ambiguous names but only unambiguous given names and surnames, and *practically unambiguous* given names and surnames, that is, words which in newspaper texts are practically always names of persons, e.g., *Archer, Boas, Brooks, Clement, Fowler, Smith*. Generally, ambiguous names like *Brown, Charlotte, Dallas, Early, Elizabeth, Knight, More, Selma, Young, Washington*, etc., are not on the names list.

The rule for identifying names of persons which are not preceded by titles states further that names in newspaper texts which are not preceded by titles contain among their elements at least one unambiguous name, for instance, *Robert* in *Robert Green*; or a word which, although ambiguous, is practically always a personal name, for instance, *Boas* in *June Boas*.

As a rule, words which adjoin unambiguous and practically unambiguous names and whose first letters are capitalized are also personal names, for instance *Green* in *Robert Green, June* in *June Boas, Mobutu* in *Joseph Mobutu*.

From this, we infer that to recognize name strings in newspaper articles, it is generally sufficient to spot unambiguous names since as a rule capitalized words and initials which adjoin them are also names and members of the same string.

Our current rule for identifying ambiguous names states that isolated occurrences of words which are both names of places and names of persons, e.g., *Alberta, Berkeley, Charlotte, Georgia, Selma, Washington* are place names.

We also assume that full names not preceded by titles and composed of ambiguous elements, e.g., *Selma Young,* and ambiguous names, e.g., *Baker, Beard, Charlotte, Young,* which occasionally occur in a newspaper article not preceded by titles or initials and not adjacent to an unambiguous name, also occur elsewhere in the same article either preceded by personal titles, e.g., *Mr. Baker, Professor Beard, Miss Selma Young, Superintendent Young,* or adjacent to unambiguous names, e.g., *Robert Baker, Charles Beard, Charlotte Corday, Susan Young.* Ambiguous names preceded by titles or adjacent to unambiguous names are identifiable as names of persons.

We assume further that if a word in a newspaper article, e.g., *Baker, Charlotte, Ford, Young, Washington,* has been recognized as a name of a person, then its other occurrences in that article—including its isolated occurrences—are also names of persons.

As a rule, if a word or a phrase in a newspaper article has been identified as a name of a person performing a special naming function with regard to the designation of some other word or phrase, e.g., *Everest* in *Mount Everest, J. P. Dumont* in *J. P. Dumont and Company, Lenin* in *icebreaker Lenin,* then all its isolated occurrences, e.g., *Everest, J. P. Dumont, Lenin,* are not names of persons but namesakes. However, if a word or a phrase in a newspaper article is identified as belonging to two or more of the following types of constructions: (1) the name of a person, *Mr. J. P. Dumont, Mr. Washington,* (2) a namesake, *J. P. Dumont and Company, Washington Brothers,* (3) a place name, *Dumont, N. J., Washington, D. C.,* and so forth, then the ambiguity of its isolated occurrences is not resolvable by our present techniques.

Our present rule for identifying namesakes and capitalized words which are not names states that most capitalized words at the beginning of sentences which adjoin personal names without being part of them, e.g., *Suddenly* in *Suddenly Robert Green . . .*, *Encountering* in *Encountering June Boas . . .*, can be listed or computed and are therefore identifiable. Likewise, we assume that most capitalized words inside sentences which adjoin personal names without being part of them, e.g., *Monday* as in . . . *on Monday Robert Green* . . ., are also listable and therefore identifiable.

We also assume that most capitalized words which adjoin personal names without being part of them and which designate namesakes, e.g., *Mount* in *Mount Kennedy, Airport* in *La Guardia Airport,* are identifiable either by "list lookup" or by means of recognition rules. For instance, if the namesake is a geographical term, e.g., *Mount, Street, Lane, Plaza,* then that term and the adjoining name of person are identifiable as a *geographical phrase* in which the name of person acts as a proper name with regard to the designatum of the geographical term, e.g., *Mount Kennedy, Kennedy Plaza.* The most frequently occurring exceptions to this rule can be listed, e.g., *Dame Lane* (the preferred interpretation of *Dame Lane* is "a Dame named Lane," rather than "a Lane named Dame," *Gallo Plaza* (the name of the former U.N. mediator to Cyprus).

If, however, the namesake designates a type of commercial establishment, e.g., *Hotel, Lodge, Radio Repair Shop,* then that term and the adjoining name of person are identifiable as *a commercial establishment phrase* in which the name of person acts as a proper name with regard to the designatum of the namesake, e.g., *Hotel Roosevelt, Wilson Radio Repair Shop.* The most frequently occurring exception to this rule can be listed, e.g., *Ambassador Lodge.*

• Results of the Experiment

Since our identification rules were embodied in dictionary entries and flowcharts which were sufficiently detailed to permit an accurate manual execution of identification procedures, it was decided that our identification system would be tested out by hand on a sample of *The New York Times* texts.[1]

Identification procedures were applied manually to some 40,000 words of texts. Altogether 88 articles from 11 issues were selected and processed (*3*, pp. 26–45). Only news articles were included in the sample. All materials found in the special sections such as entertainment, food-fashions-family-furnishings, social events, necrology, etc., were omitted. Materials in the sample consisted of only texts of news articles; picture captions, advertisements, italicized lists of various sorts, charts and diagrams, etc., were excluded from the data.

Our 40,577-word sample contained 806 occurrences of names of persons. Of the 806 occurrences of names of persons, 46 or about 6% of the total were missed. In addition, 47 words and word strings were mistakenly identified as personal names or personal titles.

Figure of merit F for the results of this identification system was computed by means of the following formula:

$$F=\frac{C^2}{(C+M)\times T}$$

where C is the number of correct identifications, M is the number of mistaken identifications, and T is the number of names of persons in the sample (*4*).

For $T=806$, $C=746$, and $M=47$

$$F=\frac{746^2}{(746+47)\times 806}= .87$$

(Because of our scoring rules [*3*, pp. 42–45] the number of identifications and misses does not add up to the number of names of persons in the sample.)

[1] Flow charts will be available at a depository library after September 1967.

• Analysis of Major Errors

Twenty-six misses (out of a total of 46) and 30 mistaken identifications (out of a total of 47) occurred in attempted identifications of words, word stems, and word strings which perform a naming function vis-a-vis some namesake, e.g., *Grumman Aircraft Engineering Corporation.* This source of misses and false identifications would be eliminated if in the future the automatic identification system were not required to decide whether words, word stems, and word strings, e.g., *Grumman,* performing a naming function vis-a-vis some identifiable namesake, e.g., *Aircraft Engineering Corporation,* are names of persons.

We also need more effective rules for computing namesake phrases, e.g., *Aircraft Company,* and personal titles, e.g., *Fireman Apprentice,* from their respective elements, e.g., *Aircraft, Company, Fireman, Apprentice.*

In addition, we need to prevent or eliminate the errors caused by the assumption that all capitalized words occurring after ambiguous words such as *General, Justice, Major, Principal,* etc., are names of persons.

We also require more effective rules to distinguish strings of titles, e.g., *President, Secretary of State,* from titles followed by names. In addition, we need more effective rules for distributing a title among all names of persons which follow it in the text, e.g., *Senators Vance Hartke and Birch Bayh of Indiana and Eugene J. McCarthy and Walter F. Mondale of Minnesota.*

Improving the automatic identification system may require several subsidiary investigations. For instance, we may be well advised to determine the relationship—if any—between, on the one hand, the effectiveness of the system and, on the other, the length, the date, the place of origin, the subject matter, the authorship, and the type of newspaper articles processed through the system.

• Some Possible Applications

In the absence of figures on the cost of identifying names of persons by computer, the subject of the applications of computer programs capable of recognizing names of persons in newspaper texts must remain in the domain of speculation.

We would conjecture that if the speed of computation were high and its price could be kept low, and if the figure of merit could be raised to .98 or higher, then a computer program for identifying names of persons in texts would be worth incorporating into existing information retrieval systems of very large newspapers and periodicals.

However, although we have reasons to think that the figure of merit could eventually exceed .95, we have no evidence that it could ever exceed .98. It is still unknown whether a program with a figure of merit lower than .98 would be useful in information retrieval. We would surmise that it might be adequate for some purposes provided that it is sufficiently fast and cheap.

Several uses suggest themselves immediately for computer programs capable of identifying cheaply, rapidly, accurately, and exhaustively the names of persons in computer-readable newspaper texts. They seem to fall into five broad and overlapping categories:

1. *Automatic indexing.* It would be possible to list by title, column, and page all newspaper articles in which some name or names appear; such a list would constitute an index to the file—possibly automated—of old newspapers.

   It would also be possible to produce biographical sketches by listing for each name of a person the headlines of articles whose texts contain that name. This technique of producing biographical sketches lends itself readily to various refinements. For example, one could produce biographical sketches of all persons whose names are preceded by a military rank.

2. *Determining how the names of persons co-occur with one another and with other words.* An automatic system for recognizing names could be used to produce a list of headlines of the newspaper articles in which certain personal names co-occur with certain other personal names, or with certain proper nouns other than personal names, or with certain common words, classes of common words, etc. It would be possible to list all such articles by date, column, page, etc., and thus to construct another type of index to the newspaper morgue.

3. *Counting occurrences of names.* It would be possible to produce frequency counts of names of different linguistic patterns: Irish, Chinese, and so forth, and to study their patterns of occurrence. A different count could be performed for each different section of the newspaper: society, business and industry, sports, entertainment, and the rest. Or a program for automatic identification of titles and names could be used to produce references to articles in which certain personal names or titles occur with certain frequency.

4. *Tracing associations between names of persons.* One procedure for tracing associations between names of persons may consist of the following series of steps: (a) recognizing names of persons in newspaper articles X; (b) searching other articles Y for occurrences of personal names which had occurred in X; (c) recognizing all names of persons in articles Y', Y'', Y''', etc., if they contain any of the personal names found in X; (d) searching other articles Z for occurrences of personal names which had occurred in Y', Y'', Y''', etc.; (e) recognizing all names of persons in articles Z', Z'', Z''', etc., if they contain any of the personal names found in Y', Y'', Y''', etc.;

   and so forth.

   This technique lends itself easily to various refinements.

5. *Providing an automatic or semiautomatic service for answering questions.* It would be possible to list the names of persons which occur in articles in which certain key terms, classes of terms, or strings of terms occur with certain frequency. Programs such as these may perhaps be helpful in providing answers to miscellaneous questions of the "Who?" type. With the addition of routines for identifying dates, place names, street addresses, and so forth (and perhaps also for parsing sentences), a program for identifying personal titles and names in texts may conceivably be expanded into an automatic or semiautomatic service for answer-

ing some questions of "Who?" "Whose?" "Whom?" "To Whom?" "From Whom?" "By Whom?" "When?" and "Where?" types.

Programs for automatic identification of names in texts may be useful to many groups, among them: (1) documentalists, librarians, and editors concerned with extracting information from texts and with automating editorial practices, research practices, etc., (2) sociologists, political scientists, and onomasticians investigating the occurrence of names in texts, (3) opinion survey and market research statisticians concerned with the occurrence of names in texts, celebrity ratings, measurement of opinion trends, etc.

• Discussion and Interpretation

Automatic classification of words and phrases of the type described in this article can be regarded as a particularly simple case of machine translation. Our algorithm attempts to identify and label certain types of words and word strings and to erase all others. In other words, the goal of this MT algorithm is text reduction: certain words and word strings are identified as "pertinent" and others as "not pertinent"; pertinent words and phrases are retained and labeled and all others are suppressed. Even this simple goal requires rules which are rather complex. However, because many word strings which the algorithms such as this one attempt to recognize have simple structure ("phrase structure"), they can be generated and possibly recognized with a reasonable degree of accuracy by a combination of linguistic and statistical techniques.

It seems that while researching and developing methods for *automatic* classification, it might be wise to set up efficient methods for *manual* classification of words in texts. In addition, it may be advisable to write programs for *automatic* extraction of data from texts which were recognized manually. Such action seems appropriate since *automatic* extraction of information from *manually* recognized texts would probably constitute a valuable service, and if and when automatic procedures for identifying dates, personal names, personal titles, trade names, company names, numbers and measure words, chemical formulas, etc., etc., become competitive with manual ones, the data processing profession will be in possession of operational computer programs capable of extracting information from recognized texts.[2]

It would seem that manual identification and classification of words and phrases in texts could be made to work efficiently and might provide a valuable interim service while methods for automatic identification are researched and developed.

Of course, the future—or to be more precise—the long-term future seems to rest with the automatic identification of words and phrases in computer-legible texts; however, more or less elaborate manual identification systems may have their moment's shine now or soon. One such procedure would require visual displays, lightpens or cursors, and computer-legible texts (*3*, pp. 55–56). Another one may look approximately as follows:

Clerks would scan printed texts (newspapers, books, journals, letters, memos, etc.) and locate various types of words and word strings (dates, names, etc.).

Upon identifying a type of word or word string, a clerk would underline it and also tag it by means of some identifying symbol. Next, key punchers would transfer to punch cards both the tags and the words and phrases identified by tags. Finally, words and tags would be transferred from cards to a suitable computer memory. Words and tags in large computer memories could then be processed by means of various cross-filing and tabulating programs. Needless to say, at this time and for some time to come, the great efficiency of a computer will be in cross-filing and tabulating.

Other solutions of this type, of KWIC type, and of related types could also be tested.

An important problem whose solution could and perhaps should be attempted now is the conversion of teletype, typesetting, typewriter, and other keyed inputs to computer-legible form. It seems that much could be done about this difficult but nevertheless resolvable problem at the base (that is, at the publishing level) and that considerable technological advance toward text processing could reasonably be expected from an intelligent cooperation of the interested parties (documentalists, publishers, administrators, hardware and software specialists, linguists, and others.[3] What the prospects for such cooperation are we do not know, but we hope that something can and will be done to make it a reality.

If this investigation of written language which mixes syntax, semantics, and pragmatics has produced some interesting observations and leads to interesting practical results, this may constitute a case for other resolutely empirical and problem-oriented investigations of texts which (1) stress formulas and results, (2) shun gratuitous axiomatization, and (3) in which the study of written language is not factored out into subdisciplines but is their product.

[2] More generally, natural language can be viewed as *macro*-language composed in part or in whole of various special-purpose *micro*-languages —each with its own structure which relative to the *total* structure of language is quite simple. It may be of some practical and theoretical interest to investigate (a) the grammars of various special-purpose micro-languages within natural language, e.g., personal titles, personal names, dates, various sets of technical and scientific terms, street addresses, trade names, place names, and (b) their structural and functional interrelations; and to research and develop automatic procedures for assigning some words and word strings in computer-readable texts to appropriate special-purpose natural and artificial micro-languages (the boundary between the two is not clearly drawn; artificial languages shade into natural).

[3] This was pointed out to me by Foster Mohrhardt in a personal communication.

References

1. Artandi, S., *Book Indexing by Computer,* Rutgers—The State University, 1963, 211 pp. Doctoral dissertation.
2. Artandi, S., Mechanical Indexing of Proper Nouns, *Journal of Documentation,* 19 (No. 4): 187–196 (1963).
3. Borkowski, C. G., *A System for Automatic Recognition of Personal Names in Newspaper Texts,* Report RC–1563, Watson IBM Research Center, Yorktown Heights, N. Y., 1966, 62 pp.
4. Meetham, A. R., Preliminary Studies for Machine Generated Vocabularies, *Language and Speech,* 6 (Part 1): 22–36 (January–March 1963).

is apparent that such coding would facilitate the retrieval and dissemination of information associated with broad classes of fiction. In the case of a patron who liked all mysteries except those by Earl Stanley Gardner, a computer program could be written to search for "all 'M' not 'ESG,'" for example.

Once written, the search logic corresponding to the interest profiles could be used either for retrospective searches or for searches of recent acquisitions. In the former case, this method would provide a bibliography of the library's holdings in the area(s) of interest; in the latter case, the method would be providing a current awareness or "alerting" service. For the retrospective searches it would seem reasonable to write a unique profile for each patron; however, for current awareness searches it might be better to economize by writing one profile for several patrons with similar interests. In any case, current awareness searches should be batched and run at intervals of time dependent on the rate of acquisition for that particular library. If the library has a high rate of acquisition, the searches should be run more frequently than if the rate of acquisition is low.

To test the feasibility of this method, IBM cards were obtained from the public library system in Lake County, Indiana. This particular library system does not have access to a computer, but it does have unit-record equipment with which it produces book catalogs. There are a little over 100 cards in the sample, and they were selected so that half are fiction and half are non-fiction. In addition, the non-fiction entries were chosen so that a variety of subject areas (and therefore DDC numbers) would be represented. A sample, one entry per line, can be seen in Fig. 1. The data on each card are arranged by fields. Starting at the left, one finds the classification number (minus the usual decimal point), the author's surname, the author's initial(s), a title or an abbreviated title, an internal code showing which branch has the item, an abbreviation for the publisher, the last two digits of the year of publication, and the price of the item.

The programs were run on the IBM 7040 computer at the Indiana University School of Medicine, and COMIT II was used as the programming language. COMIT II, a second generation list-processing language, is especially well suited for handling symbolic alphanumeric data, and its use in information retrieval work should be investigated more thoroughly in the future, particularly since it has been used almost exclusively in the area of computational linguistics (*1*).

Figure 2 shows a flowchart of the program which searches for non-fiction entries. Figure 3 shows the results of a COMIT program which will search the data base for items concerned with "the arts" and "the geography of modern Europe." Since cards are being used for input, the first action is to read the contents of the first card into workspace. If a card is found, the program has the computer check the classification number. If the classification number on the card matches the number which the computer has been instructed to find, then the entire contents of the card are printed out, and the next card is read into workspace. If the classification number does not match, the information is deleted, and the next data card is read into workspace and examined. When no card is found, the program is terminated. In the example shown, provision should have been made to ensure that the "7" and the "914" are associated only with the classification number. This can be done simply by using a null constituent to find the left end of workspace or else by revising and improving the format so that no other numerals are immediately preceded by spaces.

Figure 4 shows a flowchart of the program which searches for fiction entries. The results of a COMIT program corresponding to this flowchart are shown in Fig. 5. Specifically, this program is designed to search the data base for all westerns (which are coded "W" in this collection) and all the works in the collection by the authors Druon, Hamsun, and Spark. In a large collection it would be necessary to identify the authors more precisely, but using only the last name will illustrate the principle. The first rule which operates after a card has been read into workspace searches for an initial "W." If there is a match, the entire contents of the card are printed out, and the next card is read into workspace. If there is no match, the rule fails and control goes to a rule which finds the author's name. At this point in the program, all information other than the author's name is shelved (i.e., stored temporarily in a different place in core memory), and an internal "dictionary" is searched for a match with the author's name. If a match occurs, the contents of the shelves are called for, put in their proper positions relative to the name of the author, and the entire set of data is printed out. If no match occurs in the dictionary, the contents of the shelves are called for and are deleted along with the undesired name. Control then goes back to the first rule, and another card is found or else the program is terminated.

Since COMIT II operates with an interpreter as well as a compiler, a listing of the program rules and a terminal dump are obtained when each program is run. In addition, notification of a successful or bad compilation is provided. Both programs described here took about 30 seconds to process and run, and each cost slightly more than a dollar. Since time and cost do not increase linearly with the amount of data, these programs apparently represent an economically feasible method for retrieving this kind of information. It should also be noted that the time and cost figures include the time required to provide the listing and the terminal dump. By suppressing these two operations and by using magnetic tape for input, one should be able to make this retrieval system even less expensive.

A program of the type described in this article can be adapted to a variety of situations. The choice of the material to be disseminated to patrons should be determined by each librarian in light of the needs of his community (*2*). Moreover, it seems apparent that the use

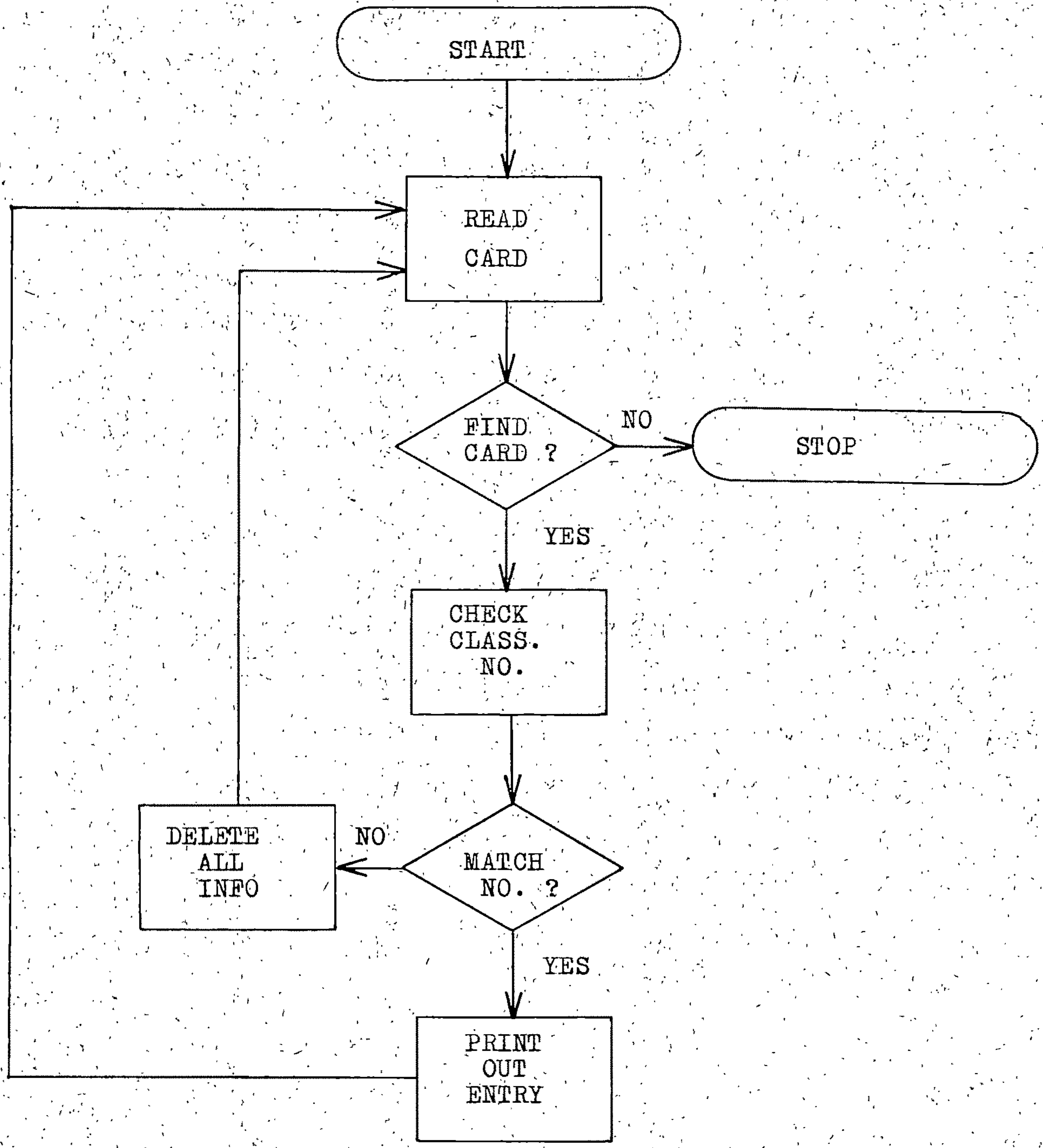
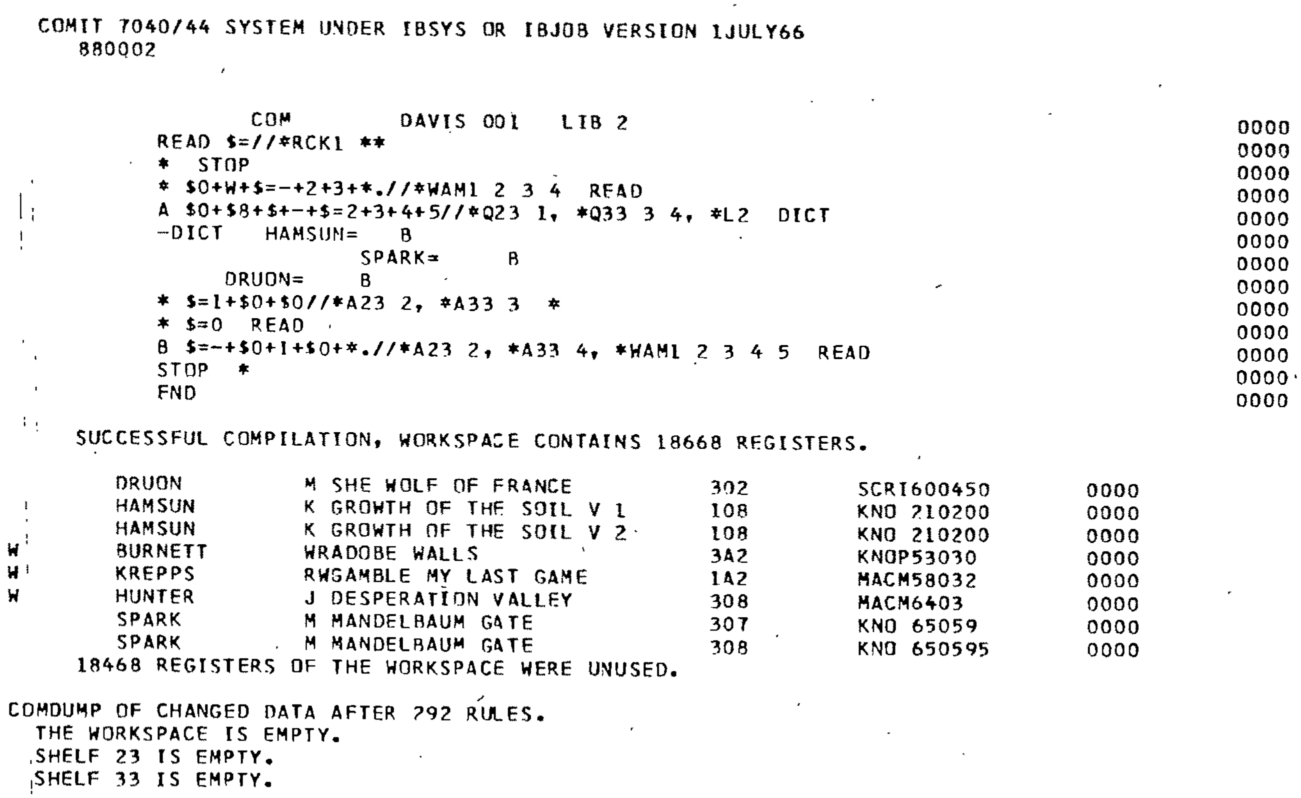


FIG. 2. Flow diagram of non-fiction search



```
COMIT 7040/44 SYSTEM UNDER IBSYS OR IBJOB VERSION 1JULY66
   880002


              COM          DAVIS 001   LIB 2                                          0000
         READ $=//*RCK1 **                                                            0000
         *  STOP                                                                      0000
         * $0+W+$=-+2+3+*.//*WAM1 2 3 4  READ                                         0000
         A $0+$8+$+-+$=2+3+4+5//*Q23 1, *Q33 3 4, *L2  DICT                           0000
         -DICT   HAMSUN=   B                                                          0000
                        SPARK=     B                                                  0000
             DRUON=    B                                                              0000
         * $=1+$0+$0//*A23 2, *A33 3  *                                               0000
         * $=0  READ                                                                  0000
         B $=-+$0+1+$0+*.//*A23 2, *A33 4, *WAM1 2 3 4 5  READ                        0000
         STOP  *                                                                      0000
         FND                                                                          0000

   SUCCESSFUL COMPILATION, WORKSPACE CONTAINS 18668 REGISTERS.

      DRUON          M SHE WOLF OF FRANCE        302       SCRI600450      0000
      HAMSUN         K GROWTH OF THE SOIL V 1    108       KNO 210200      0000
      HAMSUN         K GROWTH OF THE SOIL V 2    108       KNO 210200      0000
W     BURNETT        WRADOBE WALLS               3A2       KNOP53030       0000
W     KREPPS         RWGAMBLE MY LAST GAME       1A2       MACM58032       0000
W     HUNTER         J DESPERATION VALLEY        308       MACM6403        0000
      SPARK          M MANDELBAUM GATE           307       KNO 65059       0000
      SPARK          M MANDELBAUM GATE           308       KNO 650595      0000
   18468 REGISTERS OF THE WORKSPACE WERE UNUSED.

COMDUMP OF CHANGED DATA AFTER 292 RULES.
  THE WORKSPACE IS EMPTY.
  SHELF 23 IS EMPTY.
  SHELF 33 IS EMPTY.
```

FIG. 5. Representative fiction search

of technical services to support reader services is an area which public libraries can and should continue to explore actively.

• Acknowledgments

I wish to thank Dr. Robert Wall of the Linguistics Department, Indiana University, and Dr. Ann F. Painter of the Graduate Library School, Indiana University, for their encouragement and advice. Computations for this project were performed at the Indiana University School of Medicine Research Computation Center, which is supported in part by Public Health Service Research Grant Fr 00162-03.

References

1. M.I.T., The Research Laboratory of Electronics and the Computation Center, *An Introduction to COMIT Programming* and *COMIT Programmers Reference Manual,* The M.I.T. Press, November 1963.
2. LYMAN, HELEN H., Reader's Guidance Service in a Small Public Library, *The Small Public Library:* A Series of Guides for the Community Librarian and Trustee, American Library Association, No. 8, pp. 1–8 (1962).

The Japan Information Center of Science and Technology (JICST): Its Organization and Function

The Japan Information Center of Science and Technology (JICST) was established in 1957, with initial funds from the Japanese Government and private industry. The organization now has a full-time staff of nearly 260 people of which 25 percent consists of subject specialists. In addition, there are more than 2,000 outside cooperators for abstracting and translating. JICST's services include current bibliographic publications, photocopy, current content-sheet, translation, and literature search. The charges for these services give partial support for the activities of JICST. "JEIPAC" is a special purpose electronic computer designed for JICST's information handling, which was installed in 1961. JICST has been engaged in developing information retrieval systems in several subject areas by using this machine and it is being used in practice for metallurgical literature search. Although JICST is Japan's central organization for the dissemination of scientific and technical information, its services do not cover the fields of the life sciences because of economic limitations. The present services of JICST are mainly concerned with foreign literature relating to the physical sciences.

TAKAO FUKUDOME †

Reference Librarian
Kitasato Memorial Medical Library,
Keio University School of Medicine,
Tokyo, Japan

• Establishment and Purpose

The Japan Information Center of Science and Technology (JICST) was established in August 1957 according to the JICST Act (Law 84) and defined as "a central organization in the country for scientific and technical information." It is a nonprofit institution founded upon government and industrial contributions of 80 million yen (about $222,220; 40 million yen from the government and another 40 million from industry, i.e., firms as shareholders). The idea for such an organization had been developed since the Science and Technics Agency was established in the Japanese government in 1956. For the development of research and study in science and technology in the country, information control and handling have been a major problem inasmuch as Japan is geographically remote from the Western countries and language problems are a barrier to scientific communication.

The Science and Technics Agency started to study the problems from the viewpoint of national interest and asked the government and industries to support a new enterprise for scientific and technical information. The effort resulted in the "Japan Information Center of Science and Technology Act," passed by the Diet (Congress) in April 1957. In order to promote the development of science and technology in the country, the function of JICST is defined as follows:

1. To collect both domestic and foreign information in the field of science and technology;
2. To classify, organize, and retain that information;
3. To disseminate that information to its clientele quickly;
4. To solve problems of information handling that individual institutes or enterprises are not able to manage.

• Organization and Staff

JICST is controlled by the Prime Minister through the Science and Technics Agency, Japanese government. The president and auditor(s) are nominated by the Prime

† Now doing graduate study at Columbia University School of Library Service as a fellow of the China Medical Board of New York.

Minister, and the vice-president and director(s) are nominated by the president and with the consent of the Prime Minister, under the JICST Act, Article 13. There are now 13 advisors and 40 councilors who are representatives of learned societies, universities, and industry in the country(*1*). The principal chart of the organization is shown in Chart 1.

CHART 1. Organization of the Japan Information Center of Science and Technology

	General Affairs Division	
	Planning Office	
Advisors	Information Division	
Councilors	Documents Division	Retrieval Section Library
President		
Vice-President	Service Division	Publication Section Photoduplication Section Translation Section Investigation Section
Director(s)		
Auditor(s)		
	Osaka Branch	
	Nagoya Branch	

The Information Division is mainly concerned with editing *Current Bibliography on Science and Technology* (Kagaku Gijutsu Bunken Sokuho) (vid. IV-A). The Retrieval Section, Document Division, is engaged in indexing *Current Bibliography*, producing "information cards" (vid. IV-B), and experimenting with machine search methods. The investigation section of the Service Division offers literature search, patent literature search, and abstracting service on request from outside users who pay service fees (vid. V).

In the first year, 1957, JICST started its operation with 62 staff members; the number has been expanded to 258 (*1*, p. 2). More than 25 percent of the total staff have subject specialties, with at least baccalaureate degrees in science or technology. (In 1963, 90 of 230 staff members were subject specialists[*2*].) In addition, there are now about 2,200 outside cooperators for abstracting and translating. JICST also has two service branches in Osaka and Nagoya which are both important industrial areas in Japan.

• Acquisition and Collection of Materials

Foreign (outside Japan) current journals and patent specifications are the most important information sources in JICST. The Center received 4,135 current foreign journal titles and 1,729 domestic journals in 1964, and the number of foreign titles was predicted to increase to an estimated 4,300 in 1965 (*1*, p. 4). About 1,000 titles of foreign journals are obtained weekly by air-cargo through agents in Dusseldorf, West Germany, and in New York.

Distribution of the foreign current journals by countries and by subjects are shown in Table 1 (*3*).

Patent specifications are obtained by airfreight from the United States, England, and West Germany, and they are limited to the subject of chemistry only (vid. IV-B). About 25,500 items of the patent literature were acquired in 1964, and 27,000 items were estimated for 1965 (*3*, p. 5). The collection of books and monographs is not as large in number. It consisted of 5,400 foreign and 3,940 domestic titles at the end of March 1964 (*3*).

• Publications

A. *Current Bibliography on Science and Technology* (Kagaku Gijutsu Bunken Sokuho)

Current Bibliography is an abstracting journal that started publication in March 1958. It is now divided into ten series covering more than 4,000 foreign current journals. It contained a total of about 300,000 items in 1964 (*3*, p. 6).

1. "Chemistry & Chemical Industry series." March 1958– 3/m. (Including foreign periodicals only: 94,700 items in 1964.) Compared with *Chemical Abstracts*, the series places greater emphasis on technical interpretation of new products in the field of chemical industry.
2. "General Engineering & Mechanical Engineering series." March 1958– s-m. (Including foreign periodicals only: 46,500 items in 1964.) All articles in such important publications in this field as ASME and SAE publications are abstracted, and the main classes in the series are arranged according to the conventional classification of industries for the convenience of subscribers.

TABLE 1

1. Country	
The United States	32%
Great Britain	18%
Germany	15%
France	8%
U.S.S.R.	6%
Italy	3%
Others	18%
	100%
2. Subject *	
Chemistry and chemical industry	27%
Electrical engineering	15%
Mechanical engineering	15%
Civil engineering and architecture	8%
Geology, mining, and metallurgy	7%
Pure and applied physics	6%
Atomic energy	6%
Others	16%
	100%

* Biological sciences are not covered by the JICST services.

3. "Electrical Engineering series." March 1958– s-m. (Including foreign periodicals only: 28,500 items in 1964.)

4. "Geology, Mining, and Metallurgy series." Sept. 1958– s-m. (Including foreign periodicals only: 28,200 items in 1964.) This series especially emphasizes Soviet literature, which covers 30 or 40 percent of the total items in the series. The abstracts of Soviet literature are also longer and more detailed than those of other literatures because of language problems.

5. "Civil Engineering & Architecture series." Sept. 1958– s-m. (Including foreign periodicals only: 18,600 items in 1964.) In addition to 250 professional journals in this field, about 600 other journals received in JICST are screened for this series.

6. "Pure and Applied Physics series." April 1956– m. (Including foreign periodicals only: 30,200 items in 1964.)

7. "Atomic Energy—Radioisotope and Radiation Application series." April 1961– m. (Including both Japanese and foreign periodicals and reports: 10,000 items in 1964.) The abstracts in this series are usually longer than those in the other series; and tables, charts, or formulas are attached, if necessary. Articles in the field of nuclear medicine are not included. Japanese articles are screened from about 1,000 titles received in JICST.

8. "Business Management series." April 1963– m. (Including both Japanese and foreign publications: 12,300 items in 1964.) This series covers such subject areas as operations research, systems engineering, industrial engineering, human engineering, quality control, etc. It does not include such topics as economic statistics and federal or state economics.

9. "Chemistry in Japan—*Japanese Chemical Abstracts.*" January 1964– m. (Including 28,900 items in 1964.) About 500 titles of the 1,000 Japanese journals screened for the series are publications specializing in the chemical field. The 500 journals and Japanese patent literature in chemistry are extensively abstracted in this series.

The *Japanese Chemical Abstracts* (Nippon Kagaku Soran) itself had been published by the Japan Chemical Research Association since 1877, and was absorbed as a series of *Current Bibliography* by JICST in 1964. During 1958–1959, JICST had already published the *General Index to the Japanese Chemical Abstracts 1941–1955* (Nippon Kagaku Soran Sosakuin), which is a 15-year author and subject index to the abstracting journal.

10. "Foreign Technical Information for Smaller Enterprises series." Sept. 1965– 3/y. One of the real problems in Japanese economic and industrial development is to modernize and orient the management of medium- or small-sized enterprises which constitute the majority of Japanese industry. They do not need scientific or highly specialized research information, but they do need more practical information directly concerned with their own products. This series is trying to find new customers for the Center's services.

In addition to the abstracts, each series, except series 7 and series 9, has a "news section" in its issues. About 100 titles of periodicals of general science and technology such as *Science* and *New Scientist,* of trade journals such as *Business Week* and *Europachemie,* and of trade newspapers such as *Financial Times* and *VDI Nachrichten* are screened for this section.

B. The Compiling Process of *Current Bibliography*

The main duties of the Information Division (vid. II) are to select each article for abstracting, to check the abstracts returned from outside cooperators, and to classify them and assign UDC numbers. The Division staff consists of information specialists engaged in their own fields in the categories of the *Bibliography* series. When the three series of the *Bibliography* were first published in the spring of 1958, their contents consisted of the translated titles and one- or two-line annotations. In the autumn of that year, the annotations of the "Electrical Engineering" series were extended to "indicative" abstracts having about 150 characters. Abstracting by outside cooperators also began at the same time. In the beginning of 1959, the indicative abstracts appeared in all other series. Since 1960, "informative" abstracts have been developed with about 300 characters. Japanese is written with a mixture of Chinese characters and Japanese syllabary (KANA). In general, a Chinese character can constitute a word; therefore, a 300-character abstract can contain much more information than one having the same number of Roman characters in English. The writing system, however, creates difficulties for mechanization because of the numerous numbers of Chinese characters and their complexity.

The compiling process of *Current Bibliography* can be summarized in the following nine steps (*2, 4*):

1. Newly arrived materials are delivered from the Library to the Information Division.

2. Each article of those materials is screened for the *Bibliography* by the Division staff.

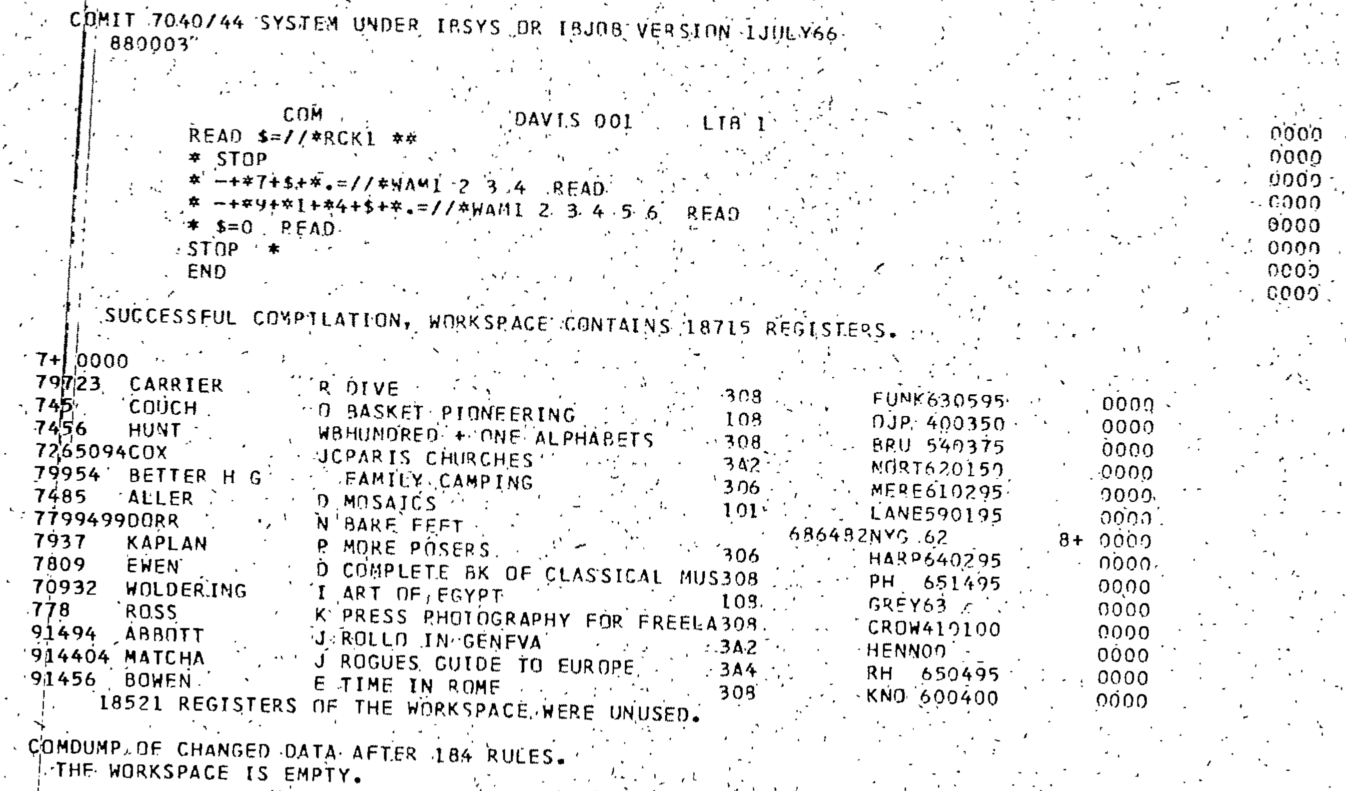
3. The selected articles are sent to outside cooperators for abstracting. The contracted abstractors and translators now number about 2,200 specialists; more than 80 percent of them work for universities or research institutes and the rest are staff connected with private enterprises.

4. The abstracts sent back from the outside cooperators are checked by the Division staff. The staff is responsible for the quality of the abstracts, uniformity of the terminology, and consistency of word usage.

5. After checking, each abstract is classified according to JICST's own classification scheme. UDC notation is also assigned, since the UDC is widely used in industrial libraries in Japan. It is said that using UDC on such a large scale of bibliographic control is comparable to that of the *Referativny Zhurnal* of the Soviet Union. At the same time, indications are made on the items which are entered under more than one category in the series.

6. The abstracts with the classification code, the UDC number, and the indication for multiple entries, if necessary, are forwarded to the Retrieval section to type "information cards"—each abstract is typed on a 4×6 card. (The abstract itself also has to be written within the limitation of the card space.) The cards for multiple entries are copied by Xerox 914 in the Photoduplication section.

7. The "information cards" are returned to the Information Division, and the duplicated entries are distributed to the appropriate series sections. Checking is done for typing mistakes and for correctness of the classification and the UDC number on the cards delivered from the other series sections.



```
COMIT 7040/44 SYSTEM UNDER IRSYS OR IBJOB VERSION 1JULY66
  880003"

                COM                    DAVIS 001     LIB 1
          READ $=//*RCK1 **                                                                   0000
          * STOP                                                                              0000
          * -+*7+$+*.=//*WAMI 2 3 4  READ                                                     0000
          * -+*9+*1+*4+$+*.=//*WAMI 2 3 4 5 6  READ                                           0000
          * $=0  READ                                                                         0000
          STOP  *                                                                             0000
          END                                                                                 0000
                                                                                              0000
   SUCCESSFUL COMPILATION, WORKSPACE CONTAINS 18715 REGISTERS.

7+ 0000
79723  CARRIER        R DIVE                          308         FUNK630595        0000
745    COUCH          O BASKET PIONEERING             108         DJP 400350        0000
7456   HUNT           WBHUNDRED + ONE ALPHABETS       308         BRU 540375        0000
7265094COX            JCPARIS CHURCHES                342         NORT620150        0000
79954  BETTER H G       FAMILY CAMPING                306         MERE610295        0000
7485   ALLER          D MOSAICS                       101         LANE590195        0000
7799499DORR           N BARE FEET                          686482NYG .62        8+ 0000
7937   KAPLAN         P MORE POSERS                   306         HARP640295        0000
7809   EWEN           D COMPLETE BK OF CLASSICAL MUS308           PH   651495       0000
70932  WOLDERING      I ART OF EGYPT                  109         GREY63            0000
778    ROSS           K PRESS PHOTOGRAPHY FOR FREELA308           CROW410100        0000
91494  ABBOTT         J ROLLO IN GENEVA               3A2         HENN00            0000
914404 MATCHA         J ROGUES GUIDE TO EUROPE        3A4         RH   650495       0000
91456  BOWEN          E TIME IN ROME                  308         KNO 600400        0000
      18521 REGISTERS OF THE WORKSPACE WERE UNUSED.

COMDUMP OF CHANGED DATA AFTER 184 RULES.
 THE WORKSPACE IS EMPTY.
```

FIG. 3. Representative non-fiction search

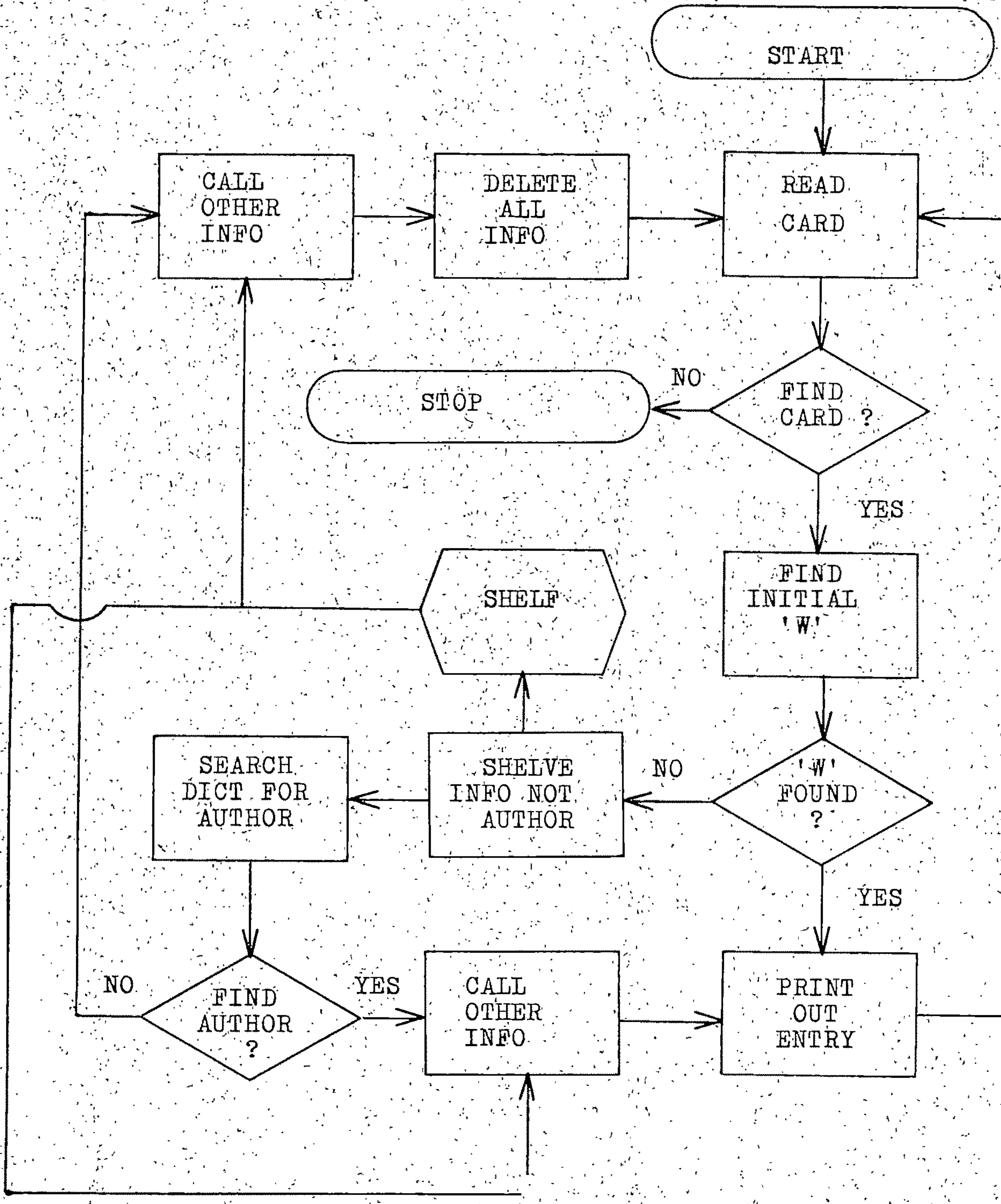


FIG. 4. Flow diagram of fiction search /

8. The finished information cards are sent back to the Retrieval section again, and those cards are filed according to classification and are assigned series item numbers.

9. The filed cards are sent to the Publication section for printing by a photo-offset method from full-card layout format.

As seen above, all the processes have been done by manual steps. However, the time-lag between the date of original publication and that of the *Bibliography* is an average of three months (*2*, *4*). After the printing, the original cards (information cards) are cumulated in the form of a card catalog which contained a total of 900,000 cards in 1964 (*2*, *4*).

C. Other publications

1. *Foreign Patent News*—Chemistry (Gaikoku Tokkyo Sokuho). April 1958– w.
This publication is compiled from the *Official Gazette* (U.S.A.), the *Official Journal* (Great Britain), and the *Patentblatt* (West Germany). It includes the following information: patent number, classification, title of the invention (translated into Japanese), applicant(s), inventor(s), application number, and application date. In the first year, 1958, the *News* was published in three editions for chemistry, electrical engineering, and mechanical engineering, but its coverage has been limited only to chemistry since 1959. The patent specifications in the three countries are obtained by air cargo. The chemistry section of the U.S. *Official Gazette* is now fully copied with the permission of the American Embassy in Tokyo.

2. *Japanese Patent Index*. (Nippon Tokkyo Sakuin) 1963– a.
This is an annual index to the *Japanese Patent Gazette*. The index consists of an "Applicant index" and a "Classification number index." The former is divided into the following three parts: "Japanese corporate names," "Japanese personal names," and "Foreign applicants." It included 19,000 patents in the 1962 edition, 26,950 in 1963, and 30,380 in 1964.

3. *JICST Monthly* (Joho Kanri). January 1958– m.
This publication was originally a house-organ of JICST, but it has now established its place as a representative professional journal of documentation in Japan. The status of the journal could be compared to that of *American Documentation* in the United States, or *Journal of Documentation* in England. Its Japanese title was changed from *Gekkan JICST* to the present one in 1963.

• Services on Demand

There are four types of services on demand: photocopies, current content-sheet service, translations, and literature searches. These direct requests from users have a close relation to *Current Bibliography* and other publications of the JICST itself. In other words, the demand services have been produced or accelerated by the publication service.

A. Photocopy Service

The photocopy service is now the heaviest business aside from the publication of *Current Bibliography* itself. In the fiscal year 1964, JICST filled more than 267,000 requests for photoduplication (*1*, p. 10). This means that about 20,000 articles or 130,000 to 140,000 pages were supplied by the service each month (*2*). It is estimated that more than 80 percent of the total requests were generated from the bibliographic publications of JICST. The requests for photocopies, therefore, have been analyzed as a tool to measure the user value of the publications. For example, a recent survey was made of the relation between photocopy requests and the "Mechanical Engineering series" of *Current Bibliography*. It is a three-month survey of 9,193 requests from March to May 1964 (*5*):

1. The items of photocopies requested in a particular time corresponded with those in a particular issue of the series; that is, the majority of the requests represented the items in an issue which had been published about 50 days before.
2. There were 70 titles which had more than 30 requests. Some of these were *Transactions of ASME, SAE Transactions, Machine Design, Metalworking Production, Design News, Machinery, Maschinenbautechnik, Mass Production,* etc.
3. English, German, and Russian, as represented by the requested literature, show a ratio of 10 : 3 : 1 whereas the languages in the series of the *Bibliography* have the ratio of 10 : 4 : 2. Hence, it can be said that Russian is not yet as popular a language in industry as English or German.
4. There were 245 articles which had more than four requests. The contents of those articles were mostly practical rather than theoretical approaches.
5. There were a great many requests for short articles, less than a page, which contained news concerning new products. This fact should be taken into consideration in the selection process of articles for *Current Bibliography*.

B. Current Content-Sheet Service

"Tables of contents" of journals requested by users are supplied in photocopies before the materials are processed for abstracting. This is a current awareness service preceding *Current Bibliography*. The service began in October, 1959, and the photocopies of tables of contents are delivered to the customers three times a month. This is not an exclusive contents service like the *Current Contents* published in the United States by the Institute for Scientific Information, but it is a selective service according to the users' requirements for particular titles.

C. Translation Service

Most of translations in the service have been done by outside cooperators who are able to contract with JICST as well as the abstracters for the *Bibliography*. Therefore, almost all foreign languages can be translated into Japanese by using the service. Translations from Japanese to foreign languages, however, are limited to English, German, French, and Russian. The requests received had risen to 5,680 in 1962, but the number has gradually decreased to 5,424 in 1963 and to 4,819 in 1964 (*1*, p. 10).

In the fiscal year 1961, JICST filled 5,022 requests for translations. The distribution by languages is shown in Table 2 (*4*, p. 16).

Fifty-four percent of the 5,022 requests were for translations of the literature abstracted in *Current Bibliography*. Others were for materials outside the *Bibliography*, including patent specifications, standards, product catalogs, correspondence, etc.

There is a great need for translations of Japanese scientific and technical literature outside Japan. For example, INSDOC in New Delhi has found some difficulty in obtaining translation service for Japanese and Chinese literature (*6*). When Dr. Mohajir, Director of PANSDOC in Karachi, visited JICST in 1962, he proposed an exchange plan for translation (*4*, p. 35). But neither institution could reach an agreement on the cost of translation. The translation service in PANSDOC has been done by the full-time staff and offered at a very small charge. On the other hand, any JICST service has to cover the expended cost because of its organizational character as a self-supported institution (vid. VII). However, there have been 30 to 50 requests a year for translations from foreign countries, and these have been provided at the regular charge.

D. Literature Search Service

Correlations have been found between the bibliographic publications of JICST, photocopy requests to JICST, and translation requests. The pattern of the JICST users' approach to information is usually limited to requests arising from the publications. However, a demand for literature search by subject may arise separately. These demands are usually generated by individual users' own production problems. In many cases, therefore, these literature searches will concern confidential matters of individual enterprises. They are afraid that their industrial securities might be violated by using the literature search service. Hence, such questions from industry frequently come in as very broad topics or are obscure in the subject contents. When there is a lack of communication between the questioner and the search staff, the result is unsatisfactory. It is important for the service to establish a reputation and to gain the confidence of its users that they may rely on the discretion of the staff in matters relating to industrial security.

TABLE 2

Translation Requests

Original—Translation Languages	
Russian—Japanese	33.5%
German—Japanese	17.8%
Japanese—English	13.0%
English—Japanese	12.8%
French—Japanese	12.3%
Others—Japanese	9.8%
Japanese—Others	0.9%
	100.00%

The literature search service of JICST does not cover the fields of medicine, agriculture, and biology. The staff and tools for the service have been well prepared for the fields of chemistry, pharmacy, electrical engineering, mechanical engineering, and metallurgy. Exclusive search for patent applications is also available from this service section of JICST.

It is expected that the literature search service will be rapidly developed when the computerization of information retrieval system in the Center comes into practice.

• Mechanization of Information Handling

There are two approaches to the mechanization of information systems in JICST. One is the mechanization of the compiling process of *Current Bibliography*, and another is the information retrieval by a computer system for the literature search. For the first purpose, IBM equipment was installed in April 1960: 24 and 26 Printing Card Punch, 56 Card Verifier, 82 Sorter, and 853 Card Type (*2*). They have been since employed for the compiling of the author index to *Current Bibliography*, *Japanese Patent Index*, periodicals holding lists, and various statistical reports. A "listcamera" is now being developed through the cooperation of JICST and Tokyo Micro Co., Ltd. The newly designed listcamera is intended to copy the subject headings and titles from the "information cards" (vid. IV-B) for compiling a subject index to *Current Bibliography*. Some disadvantages of the camera have not been solved yet: for example, cards have to be inserted into the machine by hand, so their headlines come out irregularly (*2*).

The so-called "JSIPAC" (TOSBAC 4131) is a special purpose electronic computer designed for the JICST's information handling which has been used experimentally since 1961. The processing operation includes sorting, collation, adding, subtraction, and automatic printing. The machine ability is explained in the following (*7*):

> The storage medium is magnetic tape, and four tape reels are standard. A file of documents is entered on paper tape . . . and transferred onto the magnetic tape through a high-speed photo reader. Each magnetic tape reel may include about 2,400,000 characters. One machine word contains 12 characters. Taking 10 as an average number of machine words per document, 20,000 documents can be stored on a tape reel. Tape scanning speed is 1.5 meters per second, or 9,000 characters per second. As a memory device, a 60-word magnetic core memory is installed; 25 words for the data to be processed; 15 for the question data to be searched. Then up to 20 words can be reserved for processing other than searching.

The coding required for an information retrieval system has been studied by each subject specialists' group in the Document Division. The experiment on metallurgical literature is now becoming close to a practical step.

A. The Experiment in the Metallurgy Section (*8, 9, 10, 11*)

The classification system developed for metallurgical literature by the American Society for Metals and the Special Library Association (*12*) has been studied and modified for the coding of the experiment. The experimentation by this group has been done in the following steps:

1. Sampling of key-words from the "Geology, Mining, & Metallurgy series"
2. Establishing of the coding system
3. Recording bibliographic citations
4. Punching the paper tape for transmitting the information
5. Storing the information into the magnetic tapes

The 10,392 items in volume 4 of the series already have been stored on the tapes; the 13,189 items in volume 5 were to be stored by October 1965; and the 13,025 items in volume 6 by February 1966. Then the total of 36,606 items could be employed for the literature searches in practice (*13*).

B. The Experiment in the Chemistry Section (*14*)

The most difficult problem of information retrieval in chemistry is a coding system for the chemical structures of organic compounds. The notation by the International Union of Pure and Applied Chemistry (*15*) and the information retrieval system for steroid compounds of the United States Patent Office (*16*) have been studied for the coding of chemical literature in JICST. A hand-sort punched cards system was employed as the first experimental step, and the result has been transferred into the "JAIPAC" system. The information is to be categorized according to the four facets: "Starting material," "Type of reaction," "Products," and "Object of reaction" (*2*).

C. The Experiment in the Electrical Engineering Section

The study by this group concerns the general or basic problems for information retrieval or coding with semantic coding approaches. It attempts to formalize the two semantic relations; relation between a term in the field of electrical engineering and its object, and relation between the terms in the field (*4*, p. 12). It is expected that the result would be useful not only for coding the literature in the field, but also for an automatic conversion system of symbols in general.

• Conclusion

The function of JICST is to smooth the dissemination of scientific and technical information as a central organization in Japan, and its activities have been expanded since its beginning. However, the financial status of the organization has forced it to limit its activities to some extent. JICST is not an entirely governmental institution because half of its initial fund was collected from commercial industry. Its legal status is that of a so-called "special corporation." Although it is a nonprofit institution, it is expected to expand self-support operations by its own business income. As it is impossible for JICST to operate with its business income only, annual contributions and subsidies have been given by the government since its beginning. For example, the fiscal year 1965 income budget is estimated to be nearly 800 million yen, and about half of that amount comes from government funds (*1*, p. 3).

Consequently, the services of JICST have emphasized meeting demands from industry rather than those from academic fields. There are now three major limitations to the JICST's services (*17*):

> 1. The fields of agriculture, fishery, biology, and medicine are excluded from the information handling systems because the JICST's subscribers largely consist of industrial companies whose fields of interest are in physical sciences.
> 2. The foreign patent information is limited to chemical subjects in the United States, Great Britain, and West Germany.
> 3. Japanese literature has not been included in *Current Bibliography*, except the two series of "Business Management" and "Chemistry in Japan—Japanese Chemical Abstracts."

In the field of medicine, however, the Japanese Medical Library Association (founded 1927) has developed its interlibrary loan system and the union lists of medical periodicals. As a central abstracting journal for Japanese medical literature, *Japana Centra Revuo Medicina* (Igaku Chuo Zasshi) has been privately published since 1903. A coordinated plan for the indexing of Japanese literature for the *Index Medicus* is now under discussion between the Japan Medical Library Association and the National Library of Medicine in the United States. In the field of agriculture, there is a new movement to establish a nation-wide organization which is supposed to be called "Japanese Agricultural Library Association." But its activities in the field are not yet clear.

JICST is legally under the control of the Science and Technics Agency to the degree that its funds are received from the government, and it is functionally coordinated with the Japanese Patent Office, the National Diet (Congress) Library, and the Science Council of Japan. The distinctions between the Science and Technology Division, National Diet Library, and the JICST are not always clear to the public in terms of their acquisition policies or their bibliographic organization activities as national institutions. Internationally, JICST is an associate member of FID, and the FID 1967 convention will be held in Tokyo under the auspices of JICST.

A new building for JICST is being built, and it will be completed by May, 1966. The new building was designed with the estimation that the current periodicals received in JICST would reach 10,000 titles and the annual abstracting would surpass 450,000 items in the near future. But the stack space has eight years' capacity for the collections, because it is estimated that literature more than eight years' old would become less used material in JICST (18).

• Acknowledgment

I wish to express my gratitude to Dr. Theodore C. Hines, Columbia University School of Library Science, for his invaluable advice and assistance.

References

1. *Gyomu no Gaiyo* (The business guide), Japan Information Center of Science and Technology, Tokyo, April 1965, p. 1.
2. Kitamura, An observation report of JICST, *Kagaku Gijutsu Service*, 9:14–15 (1964).
3. *Yoran 1965* (JICST guide bulletin), Japan Information Center of Science and Technology, Tokyo, 1965, p. 4.
4. *Joho Center no Ayumi* (The first five years of the JICST), Japan Information Center of Science and Technology, Tokyo, 1962, pp. 7–9.
5. Matsuo, E., A survey of the photocopy requests in JICST, No. 1, *Joho Kanri* (JICST Monthly), 8 (No. 2):38–41 (1965).
6. Krishan, A., Indian National Scientific Documentation Center, *Library Herald*, 6:32–36 (April 1963).
7. Niwa, Y., Present state of mechanization of documentation in Japan, *Revue internationale de la documention*, 29:63–65 (May 1962).
8. Abe, K., and H. Mihashi, Machine retrieval of metallurgical literature I, II, *Gekkan JICST* (JICST Monthly), 5 (No. 6):27–34, 39–43 (1962); 5 (No. 7): 24–34 (1962).
9. Abe, K., Machine retrieval of metallurgical literature III, *Joho Kanri* (JICST Monthly), 6 (No. 10):11–16 (1963).
10. Abe, K., Mechanical retrieval system for metallurgy, *Proceedings, Third annual meeting of the Committee for International Cooperation in Information Retrieval among Examining Patent Offices, ICIREPAT, Vienna, September 1963*, Spartan, Baltimore, 1964, pp. 140–147.
11. Abe, K., and I. Tominaga, Machine retrieval of metallurgical literature V: Indexing depth and retrieval efficiency, *Proceedings, Dai 2-kai Documentation Kenkyu Shukai Ronbunshu*, JICST, Tokyo, 1965, pp. 259–266.
12. *ASM-SLA metallurgical literature classification*, 2d ed., American Society for Metals, Committee on Literature Classification, Cleveland, 1958, 74 p.
13. Tominaga, I., and K. Abe, Machine retrieval of metallurgical literature VI: Literature searching for metallurgical researchers, *Proceedings, Dai 2-kai Documentation Kenkyu Shukai Ronbunshu*, JICST, Tokyo, 1965, pp. 266–273.
14. Kikuchi, T., Mechanized retrieval system for polymers, *Proceedings, Third annual meeting of the Committee for International Cooperation in Information Retrieval among Examining Patent Offices, ICIREPAT, Vienna, September 1963*, Spartan, Baltimore, 1964, pp. 148–162.
15. International Union of Pure and Applied Chemistry, Commission of Codification, Ciphering and Punched Card Techniques, *Rule for IUPAC notation for organic compounds*, Wiley, New York, 1961, 107 p.
16. Frome, J., and J. Leibowitz, A punched card system for searching steroid compounds, *Patent Office R&D Report No. 7*, July 1957.
17. Niwa, Y., Current trends of Information control and the duties of JICST *Joho Kanri*, (JICST Monthly), 7 (No. 1):3–4 (1964).
18. The outline of JICST's new building, *Joho Kanri* (JICST Monthly), 8 (No. 3):24–29 (1965).

A Framework for Comparing Term Association Measures*

The problem of choosing an association measure for some particular application is reviewed. Two methods are presented for treating various measures in a common framework—a parameterized model and a graphical interpretation of the measures. Some association measures which have been suggested to date are discussed in terms of this framework, and an example is chosen from the NASA vocabulary. Qualitative features of the list of associated terms are related to properties of the measure used, and it is suggested that this characterization can be useful in choosing which measure to use.

PAUL E. JONES and ROBERT M. CURTICE

Arthur D. Little, Inc.
Acorn Park
Cambridge, Massachusetts

Introduction

Many statistical formulas, almost all of them based on the single theoretical foundation offered by the 2×2 contingency table, have been introduced as candidates for measuring the degree, *Aab*, of association between two index terms *a* and *b*. Generally speaking, no firm relationship between the choice of association measure and its effect upon evaluated performance has been established. In part this has occurred because no opportunity for large scale in-use "tuning" of the measures to the specific needs of an operational retrieval system has been pursued to an experimental conclusion. The study of these measures has tended to stay in a research context, and few practice-oriented appraisals have been made. Nevertheless, a few comparative side-by-side appraisals of the effect of using different formulas have been conducted in pilot experiments.[1] While clear-cut preference for one formula over another (because it is a better discriminator of terms judged to be related) has not emerged from the experimental tests so far reported, the insight and experience that has been gained in laboratory tests has been valuable.

Not surprisingly, each formula has been found to have some attributes and some deficiencies. Apparently, each formula does provide, in practice, a set of associated terms among which there are many "reasonable" ones. What is annoying is that no clear-cut criterion for choice among the alternates has emerged. As a result, few candidate measures have been permanently dismissed from consideration, and a rather large set of formulas remains available.

The argument behind a typical association measure, when developed along statistical or theoretical lines, offers little or no basis for distinguishing among them. The reasoning suggests comparing the number of observed co-occurrences with the calculated number of expected co-occurrences. Given two formulas, it will generally be found that substantially this *same* supporting rationale is proffered for both; there are many ways of measuring statistical surprise or the unexpectedness of an observation,[2] and a large number of the available formulas can responsibly claim to do so. Figure 1 exhibits some of the more familiar measures and records their theoretical interpretation or rationale.

The fact that a large number of formulas has apparently survived the efforts of critical researchers to select among them is a curious problem which faces the serious student of associative retrieval. There definitely are differences in how various formulas behave. But choosing which is "best," even under stated conditions, is a problem which has only rarely been approached. In this paper we develop some of the tools we found helpful for comparing ranked term listings (profiles) produced by the use of term association measures, in practice.

* The research reported in this paper was sponsored in part by NASA under Contract No. NASW-1051, and in part by the Air Force Systems Command, Electronic Systems Division, Decision Sciences Laboratory, under Contract No. AF 19(628)-3311 with Arthur D. Little, Inc. The paper is identified as ESD-TR-67-231. Reproduction in whole or in part is permitted for any purpose of the U. S. Government.

[1] See, for example, Kuhns (*1*) and Dennis (*2*).

[2] Goodman and Kruskal (*3*, *4*) develop and elaborate this point.

If a and b are index terms, tallies of numbers of documents indexed (or not indexed) by the combinations of a and b are revealed in the 2 x 2 contingency table:

	a	not a	Total
b	fab	fb − fab	fb
not b	fa − fab	N − fa − fb + fab	N − fb
Total	fa	N − fa	N

where fa and fb are the frequencies of terms a and b respectively; fab is the number of co-occurrences of a and b , and N is the collection size.

Various measures based on this table are:

(I) $Aab = \frac{fab}{fb}$ The conditional probability given that term b is assigned to a document, that term a is also assigned.

(II) $Aab = fab - \frac{fafb}{N}$ The difference between the observed number of co-occurrences and the expected number based on chance.

(III) $Aab = \mathrm{Log}_{10} \frac{\left(\left|fabN - fafb\right| - \frac{N}{2}\right)^2 N}{fafb\,(N - fa)(N - fb)}$ The chi square formula using marginal values of the 2 x 2 table and Yates correction for small samples

(IV) $Aab = \frac{fab}{fa + fb - fab}$ The number of co-occurrences normalized by the number of documents indexed by only one of the terms.

(V) $Aab = \frac{\left(fab - \frac{fafb}{N}\right)}{\sqrt{\frac{fafb}{N}}}$ The number of standard deviations the observed co-occurrence falls to the right of the expected number of co-occurrences.

FIG. 1. The derivation of association measures based on the 2 × 2 contingency table

• Equivalent Association Measures

In practice, the use of one of the available statistical association measures serves two purposes. The first is to *select*, for a given header term, a list of associated terms, a process typically accomplished by specifying a threshold for the measure above which terms count as "associated." The second application is to provide a quantitative measure of the degree of association between the associated terms and the header. A convenient way to portray the result of applying the association measure to the data is to rank the co-occurring terms in a printed list, displaying the terms in decreasing order of the association measure being used. The order in which the terms are presented on such a "profile" exhibits whether one term is more[3] associated with the header term than another term is.

In attempting to choose among association formulas, we postulated that our first concern was to find a measure which yields an acceptable *ranking* of associated terms. The magnitudes of the numerical values for the degree of association are initially of no interest. What matters is whether the most closely associated term under formula *A* is one of the most closely associated terms under formula *B*. If so, the formulas are similar; if not, dissimilar. Generalizing these ideas, two formulas which yield the same (or substantially the same) ranking

[3] This is true except where there are ties produced by the association measure in use.

of terms in the profile are equivalent from this point of view.

The notion of equivalent rankings is important principally because this is the practical way to tell the formulas apart. One prepares profiles using several formulas and examines them to see which one places the most suitable terms at or near the head of the list. Attention to the numerical values assigned is secondary. Since we are trying to relate the behavior of the formulas to the kinds of things a person comparing such profiles side-by-side would look for, the ordering is the property to examine first.

• Generating a Spectrum of Association Measures

The objective of comparing the ranking behavior of various formulas is well served by finding a useful way to place them all into the same mathematical form. One way to do this is to develop a general expression that generates all the formulas of interest and that reduces to any specific one by a choice of parameters in the general expression. But a glance at the expressions for *Aab* in Fig. 1 shows that a general expression that would include, for instance, formula III as a special case would be too complicated to manage. Fortunately, by directing attention to the approximate *ranking* produced by a formula, it is possible to use a simple, readily understandable model for generating a useful spectrum of alternatives. We shall treat the process of forming an association list as a stylized method of *retrieving* certain documents.

Let term *a* with frequency *fa* be the header term for which we wish to develop a profile. Let some other term *b*, with frequency *fb*, co-occur with *a fab* times, as shown by the matrix in Fig. 2. Let us now think of *a* as defining (as it clearly does) a set of documents: those documents indexed by *a*. Let us think of the other terms *b* in the vocabulary (candidates for being "associated" with *a*) as single-term requests, and define the objective of each of these searches to be the retrieval of those documents indexed by *a*. In short, the *a*-indexed documents (and only those) are "relevant." The *b* indexed documents are "retrieved."

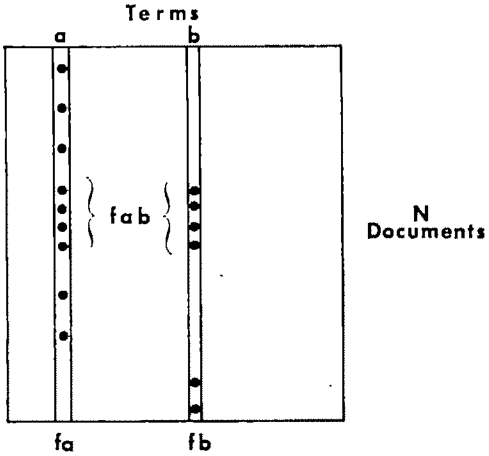


FIG. 2. Document-term matrix showing that term *a* (frequency *fa*) co-occurs with term *b* (frequency *fb*) exactly *fab* times

With this conceptual attitude, the familiar Recall and Precision measures can be defined for each term *b* (with respect to the given term *a*). They measure—with the usual disclaimers—the goodness of *b* as a substitute for *a*.

The Recall of term *b* is the proportion of documents indexed by term *a* which are also indexed by term *b*:

$$\text{Recall}_b = \frac{fab}{fa} \tag{1}$$

The precision of term *b* is the proportion of documents posted to *b* which are also indexed by term *a*:

$$\text{Precision}_b = \frac{fab}{fb} \tag{2}$$

We now have two measures of *b*'s capability to be used in lieu of *a*. But we want only one since the degree to which *b* is associated with *a* is a single number. Therefore we wish to combine Recall and Precision into a single measure. The product suggests itself since a term *b* with both high Recall and Precision should have a high association value. But since we have no idea whether to consider Recall more important than Precision or vice versa, we multiply them together with adjustable exponents. Thus a spectrum of directly interpretable measures of the association of term *b* with term *a* is provided by

$$Aab = \left(\frac{fab}{fb}\right)^{(1-n)} \left(\frac{fab}{fa}\right)^{n} \tag{3}$$

where n is such that $0 \leq n \leq 1$. Varying n generates a variety of association measures, each representing a different interpretation of the relative importance of Recall and Precision in this viewpoint towards association measures.

Since we ascribe little merit to the actual numerical value *Aab* given by a measure, putting emphasis rather on the resulting profile term ranking, we allow ourselves to alter a given measure—including this one—so long as the ranking remains invariant under the alteration. Two alterations of this kind which yield equivalent rankings are important:

a) Any positive power of a formula which yields non-negative association measures produces the same ranking of terms as the original formula.

Proof: $Axy > Axz > 0$ and $k > 0$ implies $Axy^k > Axz^k > 0$

b) Also, we will usually be allowed to strike the factor *fa* from the measure, since it is the same constant for all the terms in *a*'s profile and therefore does not affect the ranking.

Applying these rules to Equation 3 yields the equivalent formulation

$$A_{ab} = (fab)^{(1-n)}\left(\frac{fab}{fb}\right)^{n} = \frac{fab}{fb^{n}} \tag{4}$$

The model presented above thus yields a spectrum of measures of the type $Aab = fab/fb^{n}$. Each such measure is "rank-equivalent to" (produces the same ranking as) a formula derived from a specified weighting of Recall and Precision in this framework.

• Graphical Interpretation of the Rankings Produced

The next task is to find, for the more complex statistical formulas, which choice of n produces substantially the same ranking of terms. This will allow us, if we choose, to interpret those other formulas in a common framework. Let us therefore examine more closely the way the choice of n affects the ranking and develop some of the apparatus for relating n to more complex measures.

Figure 3 shows a graph of the (fb, fab) space which is of interest because of the form of Equation 4. Each term which co-occurs with the header term can be placed as a point on this graph according to its frequency (fb) and the number of times it co-occurs with the header term (fab). (Note that all terms must be located on or below 45° line, since $fb \geqq fab$.)

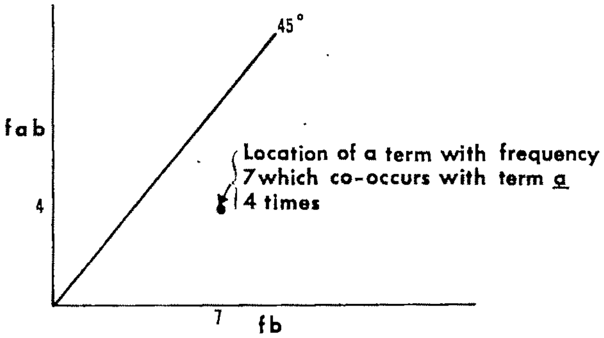

45°
fab
4
Location of a term with frequency 7 which co-occurs with term a 4 times
7
fb

FIG. 3. The (fb, fab) space

Figure 4 indicates what the distributions of fab and fb might be[4] for the terms which co-occur with a given header term. An association measure is represented dynamically as a curve of stated shape which moves in this space, and the ranking of terms on a profile according to that measure is the order in which this curve passes points which represent co-occurring terms.

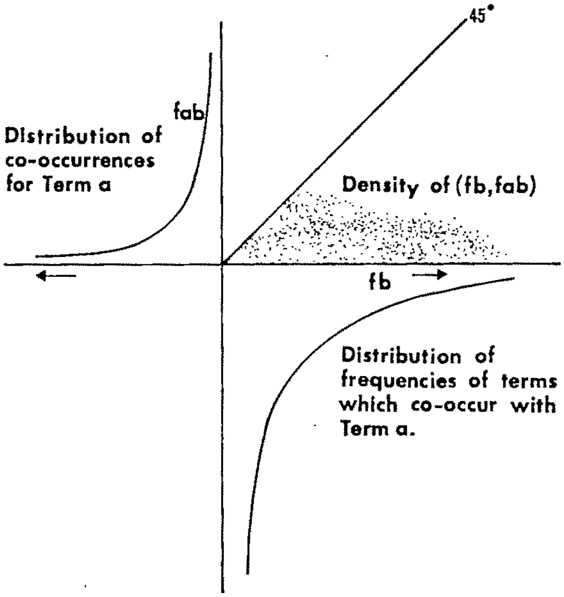

45°
fab
Distribution of co-occurrences for Term a
Density of (fb,fab)
fb
Distribution of frequencies of terms which co-occur with Term a.

FIG. 4. Distribution of points in the (fb, fab) space

Viewing the measures in this way allows us to perceive which areas of the space are passed first and therefore which terms are likely to be highly ranked.

In Fig. 5 we have shown the curves and movements for various values of n in Equation 4.

When $n=1$, we have a straight line rotating clockwise about the origin. This is the representation of the measure

$$Aab = \frac{fab}{fb}$$

which is pure Precision in terms of our model. Note that the maximum value attainable arises when $fab = fb$, i.e., term b co-occurs with term a each time it occurs. Thus, a term with the values $fb = 1$, $fab = 1$, or equivalent must be ranked number 1. No distinction is made between such a term and one with values $fb = 10$, $fab = 10$, although there seems to be more evidence in support of saying the terms are associated in this latter case.

The graphical representation of the case for $n = \frac{1}{2}$ is

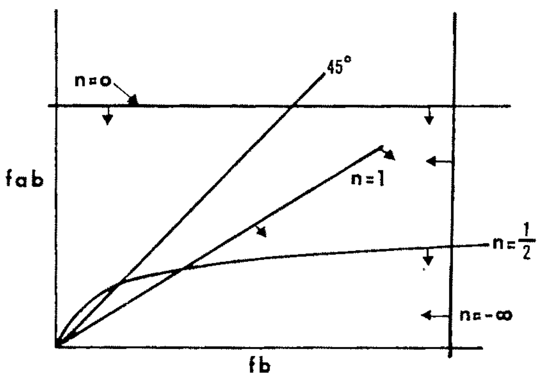


FIG. 5. Graphical representation of various association measures suggested by the model

[4] It would be possible empirically to derive the distribution of (fb, fab) by sampling a large portion of the data.

While we have the raw data, we have not determined the detailed distribution. However, it is known by observation that the points strongly tend to be distributed in the manner shown in Fig. 4, with a very high density of points in the lower left-hand corner, trailing off horizontally and upwards. For the present purposes this crude description of the distribution is probably sufficient, though more accurate data would be helpful in future work.

a curve as shown in Fig. 5 again rotating clockwise about the origin. The measure for $n=\frac{1}{2}$ is

$$Aab=\frac{fab}{fb^{\frac{1}{2}}}$$

In contrast to the case for $n=1$, this formula ranks a term with values $fb=10$, $fab=10$ higher than a term with values $fb=1$, $fab=1$. This phenomenon is quite apparent in the dynamic graph since different areas are covered first by the movement of the corresponding curves. The bend in the curve for $n=\frac{1}{2}$ causes it to be well above the point $fb=1$, $fab=1$ when it crosses the point $fb=10$, $fab=10$. In general, measures of the type fab divided by some root of fb tend to bend more sharply as n approaches 0.

The limiting case when n does reach 0 is given by a horizontal line which moves straight downward with decreasing fab. This measure ranks the terms in order of co-occurrence count. (This is pure Recall.)

Not all possible ranking rules fall within the scope of the model, of course. For example, the extreme case represented by a vertical line that moves from right to left is not strictly within the scope of the model, since n would have to be $-\infty$. It is of interest as an extreme, however, and for this reason is shown in Fig. 4. It corresponds to the measure

$$Aab=fb$$

The dynamic graphical behavior of the formulas produced by the model supplies a means for visualizing the effects which various choices of n will have on the ranking of terms. Other more complex statistical formulas can also be treated within this same framework.

We shall show in the next section that the graphical behavior of these other formulas often resembles the graphical behavior of those generated by the simple model with a suitable choice of n.

• Relationship with Other Measures

In general, the shape of a curve corresponding to some association formula outside our fab/fb^n framework can be obtained by setting the measure equal to a constant, provided the formula is a function of (fab, fb). The constant represents a particular threshold value; the movement of the curve corresponds to alteration of the threshold. As the threshold is reduced, the curve moves downward and a ranking results.

The measure given by

$$Aab=fab-\frac{fafb}{N} \qquad (5)$$

where N=collection size, has been suggested by Maron and Kuhns (5). When set to a constant (to represent a particular threshold K), and solved for fab, Equation 5 becomes

$$fab=K+\frac{fafb}{N} \qquad (6)$$

The graphical representation of fab as a function of fb is thus a straight line with small[5] slope fa/N, emerging from the vertical axis at point $fab=K$.

As the threshold is reduced, this nearly horizontal line moves vertically downward. The ranking it produces can be expected to be very similar to a ranking by fab alone, except that the slope of the line is fa/N rather than zero. Note that we cannot strike fa or N from the formula without disturbing the ranking since they are not pure additive or multiplicative constants.

A similar graphical interpretation is given for

$$Aab=\frac{fab}{fa+fb-fab} \qquad (7)$$

a measure suggested by Doyle (6).

When set to K and solved for fab we get

$$fab=\frac{K}{K+1}fb+\frac{K}{K+1}fa \qquad (8)$$

Again a linear relationship exists between fb and fab as in the previous case. However, the slope of the line approaches zero as the line moves from the top of the graph (where the slope is 1) to the bottom. (This one requires more effort than the rest to see clearly in the dynamic graph.)

In the formula suggested by Dennis (2) and given by

$$Aab=\frac{fab-\frac{fafb}{N}}{\sqrt{\frac{fafb}{N}}} \qquad (9)$$

we may eliminate fa and N as constants. Setting Aab to K and solving for fab yields

$$fab=K\sqrt{fb}+\frac{fa}{N}fb \qquad (10)$$

this is seen to be representable as the sum of the square root curve (i.e., $n=\frac{1}{2}$ and a straight line with slope fa/N. Since we would expect fa/N to be small, Equation 9 should yield a ranking similar to the case when $n=\frac{1}{2}$.

A measure investigated by Stiles (7) and given by

$$Aab=\log_{10}\frac{\left(|fabN-fafb|-\frac{N}{2}\right)^2 N}{fafb\,(N-fa)(N-fb)} \qquad (11)$$

is approximately equivalent to the ranking produced by the measure

$$\frac{(fab-\frac{1}{2})^2 N}{fafb} \qquad (12)$$

when $fa \times fb$ is less than N. (See Fossum (8).) Striking out N and fa we can see that this measure will approximate the case when $n=\frac{1}{2}$. The representation of this measure then will be a curve very similar to

$$fab=K\sqrt{fb} \qquad (13)$$

The formulas treated above (except for Doyle's) are thus convertible, without extraordinary difficulty into a form where a value of n in fab/fb^n can be assigned as a crude descriptive parameter. We do this because of

[5] Only a handful of terms a (in the NASA vocabulary, 10 out of 18,000) have fa/N in excess of .03.

a desire to compare the formulas within the simpler framework provided by the Recall and Precision model discussed earlier. The relations will be clearer after studying the illustrative examples in the next section.

• Illustrative Example

The relationships among the various formulas discussed in the preceding section are illustrated in the example presented in Fig. 6. For reasons of space and clarity, the length of the association lists is radically curtailed. The data are drawn from the NASA collection statistics, a collection we noted earlier which contains about 100,000 documents and 18,000 index terms.

Figure 6 illustrates the 15 top associates for the header term "Rocket Motor Case" as ranked by each of 9 measures. The set of co-occurring terms was first restricted to include only those terms b such that 2% or more of their occurrences were co-occurrences with "Rocket Motor Case," i.e., $fab/fb \geqq .02$. This restriction cut the set of candidates to about $\frac{1}{3}$ of the set of all terms that co-occurred with the header term, and was necessary because of computer program limitations. This selection governs all the lists compared in Fig. 6.

Five of the rankings shown are produced by arranging the term b according to five choices of n in Equation 4. The remaining 4 rankings were produced by measures in Fig. 1.

Figure 7 shows the position of the curve representing each of the measures as it selects its 15th term. That is, in the positions indicated, each of the measures has chosen 15 terms associated with "Rocket Motor Case." The thresholds corresponding to these positions vary, and for a particular measure the threshold is merely the value of that measure for the term ranked 15. Thus, by setting the measure equal to the threshold represented by the 15th ranked term, the equation of the curve at that point results.

The various term lists shown in Fig. 6 are arranged systematically according to the average slopes of the corresponding curve in Fig. 7, beginning with the vertical line (fb) and rotating counter clockwise until we reach the 45° line. This arrangement corresponds to increasing n from $-\infty$ to $+1$ for those measures derived from the model; the other measures are interspersed.

Inspection of the lists in Fig. 6 will quickly indicate that they all contain a good proportion of terms which most evaluators would judge to be associated with the header term, "Rocket Motor Case." This is not particularly surprising in a vocabulary of 18,000 terms. There are probably 150 good associates for each middle frequency term in a vocabulary of this size. Even if it were practical to print the top (say) 200 terms here, side-by-side appraisal of the comparative rankings would require more effort than the reader would expect (or want to devote). Fortunately, however, we can characterize each list by describing the types of terms which tend to appear at the top of the list. These characteristics are features of the measure which generated the particular list. Given a specific application for which the list is to be used, we could then hope to assess in advance which measures would be expected best to meet the application's requirements. Essentially, all that is needed is some statement about which characteristics of highly associated terms are desirable for the given application.

The list for $N=-\infty$ i.e., ranking by fb alone, is surprisingly good, even when we recall the 2% selection restriction mentioned above. That is, merely selecting the high frequency terms which co-occur with the header more than 2% of the time—then ranking them by decreasing frequency—yields a list of words that is far from ridiculous. This indicates, in fact, that the term co-occurrence phenomenon is a stronger effect than one might be predisposed to suspect. Naturally, this list contains terms which tend to be very general, broad, highly used vocabulary terms (by construction). It has the noteworthy attribute that there are hardly any terms appearing on this list which one needs special knowledge to understand. (The vertical line representing this measure was at $fb=637$ when the 15th term was chosen.)

The terms on the list for $n=0$, i.e., ranking by fab, are quite similar to those on the list for $n=-\infty$. Note, however, that the constituent terms "case" and "motor" have moved into prominence on this list. (The horizontal line which represents this measure had the Equation $fab=31$ when the 15th term was chosen.)

The ranking of the top 15 terms for the measure $fab-fafb/N$ is precisely that of the previous measure, fab. Thus the factor $fafb/N$ was not great enough to influence the ranking up to this point. Note the line at this point has already turned clockwise to a very small slope, the equation of the line being

$$fab = 28 + .00243fb$$

The list given by the measure suggested by Stiles (7) differs significantly from the previous lists. Technical terms, like "Hydro test," "deep draw," and "closure" begin to be included. This list and the next two are extremely similar, with only minor permutations of terms. (The curve for this measure was obtained by plotting actual points since the equation is quite complex.) However, Fig. 7 exhibits the obvious similarity of this and the next two measures, showing that the approximation in Equation 12 is indeed valid for this header term. The equations of the curves for the measures given by

$$\frac{fab-\frac{fafb}{N}}{\sqrt{\frac{fafb}{N}}} \text{ and } \sqrt{\frac{fab}{fb}}$$

$fb\ (n = -\infty)$	$fab\ (n = 0)$	$fab - \frac{fafb}{N}$
Rocket	Case	Case
Propellant	Motor	Motor
Solid	Rocket	Rocket
Steel	Rocket Engine	Rocket Engine
Rocket Engine	Solid	Solid
Fabrications	Propellant	Propellant
Titanium	Steel	Steel
Motor	Fabrication	Fabrication
Glass	Winding	Winding
Grain	Filament	Filament
Insulation	Solid Prop. Rocket Eng.	Solid Prop. Rocket Eng.
Welding	Filament Winding	Filament Winding
Fracture	Titanium	Titanium
Bonding	Fiber	Fiber
Cryogenics	Glass	Glass

STILES	$\frac{fab - \frac{fafb}{N}}{\sqrt{\frac{fafb}{N}}}$	$\frac{fab}{fb^{1/2}}\ (n = 1/2)$
Case	Case	Case
Motor	Motor	Motor
Winding	Winding	Winding
Rocket	Filament Winding	Rocket
Filament Winding	Deep Draw	Filament Winding
Stretch Forming	Rocket	Deep Draw
Deep Draw	Stretch Forming	Stretch Forming
Hydrotest	Hydrotest	Hydrotest
Filament	Filament	Filament
Rocket Engine	Rocket Engine	Rocket Engine
Closure	Closure	Closure
Fiberglass	Fiberglass	Fiberglass
Stretch	Stretch	Stretch
Steel	Spiral Wrap	Spiral Wrap
Fabrication	Steel	Steel

$\frac{fab}{fa + fb - fab}$	$\frac{fab}{fb^{2/3}}\ (n = 2/3)$	$\frac{fab}{fb}\ (n = 1)$
Case	Case	Altair Missile
Motor	Deep Draw	Polaris A2A Missile
Winding	Motor	Seepage
Filament Winding	Spiral Wrap	Spiral Wrap
Filament	Stretch Forming	Stretch Project
Closure	Hydrotest	TU 290 Motor
Fiberglass	Seepage	Turks Head Mill
Glass Fiber	Winding	Deep Draw
Stretch Forming	Filament Winding	Environmental Temp.
Stretch	Altair Missile	Fuzz
Solid Prop. Rocket Eng.	Stretch Project	Helical Winding
Rocket Engine	Closure	Stretch Forming
Reinforced Plastic	TU 290 Motor	Hydrotest
High Strength	Turks Head Mill	Vasco Jet
Fabrication	Polaris A2A Missile	Wing IV Motor

FIG. 6. The top 15 associates of the term "Rocket Motor Case" as ranked by each of nine association measures

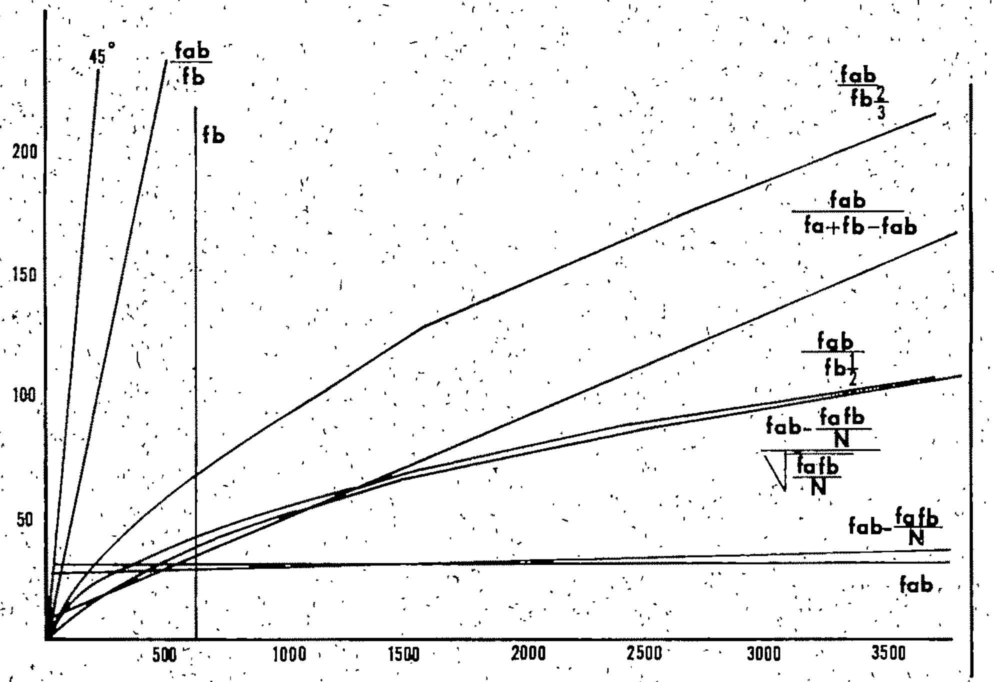


Fig. 7. Graphical representation of various association measures as they select term no. 15

are $fab=1.58\ \sqrt{fb}+.00243fb$ and $fab=1.73\ \sqrt{fb}$, respectively.

The ranking by $fab/fa+fb-fab$ contains many of the previously seen terms, plus some specific additions such as "high strength" and "reinforced plastic." The curve representing this measure is given at this point by

$$fab = .041fb + 10$$

Next, the list for $n=\frac{2}{3}$, i.e., $fab/fb^{\frac{2}{3}}$, shows the addition of some very specific, highly technical terms such as "seepage," "Polaris A2A missile" and "Turks head mill." The equation of the curve when it has chosen the 15th term is

$$fab = .9fb^{\frac{2}{3}}$$

Finally, the list given by fab/fb when $n=1$ is presented. Almost all 15 terms on the list are low frequency, highly specific terms. The equation of the line is

$$fab = .46fb$$

at this point.

The discussion above, in conjunction with an inspection of the curves representing various measures, shows that the lists corresponding to various n tend to become more specific and technical in nature as n goes from $-\infty$ to $+1$.

• Conclusion

The ranking of associated terms produced by using a particular association formula is capable of being evaluated subjectively (though crudely) for its value in a particular application. A gross parametrization of the range of possibilities to consider is presented, and the parameter n is suggested as a summary statistic. It appears to covary with the kinds of subjective characteristics of the lists that bear upon choosing a formula for a particular application. In particular as n moves toward 1 the lists become increasingly specific and specialized.

It is important to repeat that we ascribe no great theoretical importance to the parametrization in terms of n. We regard it as a convenient way to encapsulate a host of crude but interesting empirical observations. We recognize that there are workers who care about, and situations that call for, very detailed attention to the exact value of the association coefficient. We regard the present development principally as a way to narrow the field to isolate the principal behavioral features of a formula with useful properties. The exact choice, once this rough one has been made, of the particular formula best suited to the application is a matter that those closest to the situation are best qualified to study.

Given the framework developed here, we now turn in conclusion to summarizing our own experience, observations, and impressions as they relate to the spectrum of profiles.

1. As a rule one can find people and applications for which each of the lists generated (e.g., in Fig. 6) is evaluated as "best." The choice depends *both* on characteristics of the person using the list and on the nature of the intended application.

2. *People*—even technical specialists tend to reject the lists where n approaches 1 because of the prominence

given to extraordinarily technical, specific, and mysterious terms. People tend to find unfamiliar terms like "Turks Head Mill" disconcerting. However, for fully automatic associative retrieval (when a request is expanded into a profile, and that profile—without being inspected—is used as the weighted search prescription) the formulas with n near 1 appear to give the best performance.

3. For the purpose of preparing a Thesaurus listing for use by subject matter specialists either during query formulation or during an interactive retrieval process, the formulas in the vicinity of $n=\frac{1}{2}$ appear to be most satisfactory. As a rule, the more special knowledge of the field which one can count on the user possessing, the more we can move toward formulas with higher values of n. The range $\frac{1}{3} \leq n \leq \frac{2}{3}$ seems most interesting for applications of this type.

4. We have had only limited experience with the effort to use an association profile in preparing a printed Thesaurus for general use. Preliminary indications are that measures in the low range of n $(0 \leq n \leq \frac{1}{3})$ tend to be preferred.

To make a very broad summary statement, there is a tendency for n to vary more or less in accordance with "how far along in the retrieval process one is." By this we mean that a requestor, entering the retrieval situation for the first time, who is not too sure of what he is looking for or how to express it, is likely to have a preference for lists with low values of n. As he moves along through the search or gains experience with the collection, the vocabulary, the field, etc., higher values of n become appropriate. (Concurrent with these developments, of course, are considerations of whether the requestor is expanding or narrowing his search at the time he is inspecting a particular association list. Higher values of n tend to be most appropriate for narrowing.) Finally, as the requestor nears the point where he is ready to look at documents, a fairly high value of n seems to be most appropriate.

Adjusting the parameter during the course of an interactive search with the requestor on-line may be of value if this apparent trend is substantiated and if the practicalities of the system permit.

But for the present we suggest this view mainly as a bridge by which the experience, intuition, and judgment of those closely involved in a particular retrieval application can be related to the choice of an association measure.

References

1. Kuhns, J. L., The Continuum of Coefficients of Association, *Statistical Association Methods for Mechanized Documentation*, National Bureau of Standards Miscellaneous Publication 269, Washington, D. C., 1965.
2. Dennis, S. F., The Construction of a Thesaurus Automatically from a Sample of Text, *Statistical Association Methods for Mechanized Documentation*, National Bureau of Standards Miscellaneous Publication 269, Washington, D. C., 1965.
3. Goodman, L., and W. Kruskal, Measures of Association for Cross-Classifications, *Journal of American Statistical Association*, 49:732–764 (1954).
4. Goodman, L., and W. Kruskal, Measures of Association for Cross-Classifications. II: Further Discussion and References, *Journal of American Statistical Association*, 54:123–163 (1959).
5. Maron, M. E., and J. L. Kuhns, On Relevance, Probabilistic Indexing and Information Retrieval, *Journal of the Association of Computing Machinery*, 7:216–244 (1960).
6. Doyle, L. B., Indexing and Abstracting by Association, *American Documentation*, 13:378–390 (1962).
7. Stiles, H. E., The Association Factor in Information Retrieval, *Journal of the Association of Computing Machinery*, 8:271–279 (1961).
8. Fossum, E. G., et al., *Organization and Standardization of Information Retrieval Language and Systems*, UNIVAC, Blue Bell, Pa., 1966.

Cost Distribution and Analysis in Computer Storage and Retrieval. II

Additional data are presented on the cost of computer storage and retrieval activities. The effects of system modification and new hardware are noted.

H. MARRON and M. SNYDERMAN

Science Information Exchange
Smithsonian Institution

In an earlier paper (*1*), we proposed a method of allocating computer costs in a mechanized storage and retrieval activity. Also included was actual operating cost experience for the Science Information Exchange (SIE) from the period January 1964 to June 1965. System modifications and equipment changes since then have continued to reduce costs further as shown in Table 1.

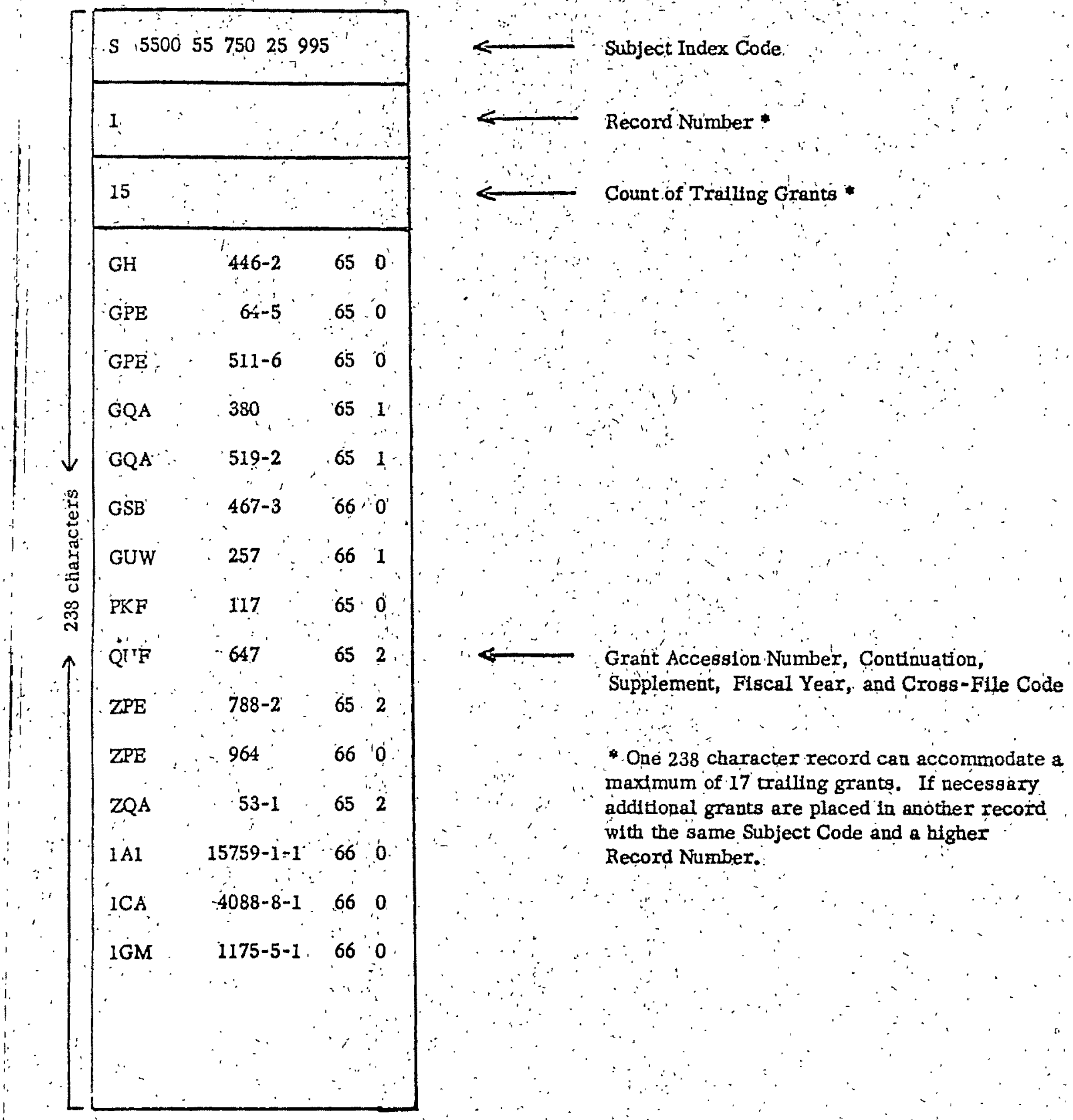
In May 1965, an IBM 1460 central processing unit replaced the Exchange's IBM 1401. In April 1966, SIE replaced its IBM 1460 array with an IBM 360/30 and added direct access disc capability to the tape oriented system. The master file was retained on magnetic tape and from it an inverted subject file was generated on discs enabling direct access searching. The inverted disc file contains a list of all the subject index points used at SIE. Appended to each point are all the identification numbers of projects which have been indexed with that particular point. This is illustrated in Fig. 1. The search for any subject category goes directly to the disc area containing the desired point and immediately supplies the results. Matches for projects containing two or more points are accomplished simply by finding which identical identification numbers appear under each.

The necessary computer programming for direct access disc searching was completed late in the summer of 1966. September 1966 was the first full month in which the inverted subject search system was used for many tasks that would have been batched and run against the magnetic tape master file. The cost reduction per job has been in accordance with expectations, and costs are expected to go even lower as further refinements are made to the operating system.

Area 1 of Table 1 shows a summary of operating experience presented in our previous paper. The batched jobs (i.e., subject or bibliographic searches which for economy considerations were batched and run against a single pass of a master tape file), declined from $37 per job in early 1964 to about $30 per job in mid-1965.

Table 1. Allocation of computer costs

						Direct computer costs						Total computer costs					
		Computer use				Batched jobs			Singly run jobs			Batched jobs			Singly run jobs		
Area	Period	Total hrs.	Total cost	$/Hr.	Maint. hrs.	Hrs.	Jobs	$/Job	Hrs.	Jobs	$/Job	Hrs.	Jobs	$/Job	Hrs.	Jobs	$/Job
1	1/64–6/64	1593	$73,900	47	384	523	629	37	703	217	152	658	629	49	928	217	202
	7/64–12/64	1752	74,450	42	386	631	617	36	846	309	116	667	617	46	1090	309	149
	1/65–6/65	1728	76,500	44	466	545	932	30	631	200	139	871	932	41	857	200	190
2	7/65–12/65	1922	83,008	43	625	545	1131	21	751	334	97	812	1131	31	1110	334	139
	1/66–6/66	1750	81,125	46	526	655	1365	22	569	173	150	945	1365	32	805	173	213
3	9/66	240	14,219	59	61	51	212	11	98	77	75	82	212	23	158	77	121



238 characters
Subject Index Code
Record Number *
Count of Trailing Grants *
Grant Accession Number, Continuation, Supplement, Fiscal Year, and Cross-File Code

S 5500 55 750 25 995			
1			
15			
GH	446-2	65	0
GPE	64-5	65	0
GPE	511-6	65	0
GQA	380	65	1
GQA	519-2	65	1
GSB	467-3	66	0
GUW	257	66	1
PKF	117	65	0
QUF	647	65	2
ZPE	788-2	65	2
ZPE	964	66	0
ZQA	53-1	65	2
1A1	15759-1-1	66	0
1CA	4088-8-1	66	0
1GM	1175-5-1	66	0

* One 238 character record can accommodate a maximum of 17 trailing grants. If necessary additional grants are placed in another record with the same Subject Code and a higher Record Number.

FIG. 1. Disc record used for subject search (Schematic—August 1966)

When a prorated share of the maintenance expense is added, however, the costs range from $49 to $41 per search over the same time periods.

Area 2 shows the results of some systems improvements plus the faster cycle time of the 1460 central processing unit. The direct cost per search was reduced to about $22 or about $32 per search with the maintenance burden added. The costs per tape search did not decrease further upon installation of the 360 until disc searches were initiated.

Area 3 shows costs for the first month when most of the searches were performed using the disc files. Master tape files still had to be used in certain special cases. The cost per subject search was further reduced to $11 per job for direct cost and $23 per job with maintenance burden added.

Figure 2 shows the cost data in graphic form. It is interesting to note that the cost which includes the maintenance burden is not as dramatically reduced as is the direct computer cost. This is partially because it requires additional maintenance hours to update the inverted files.

We expect the cost per bibliographic search to decrease further because: (1) a greater proportion of

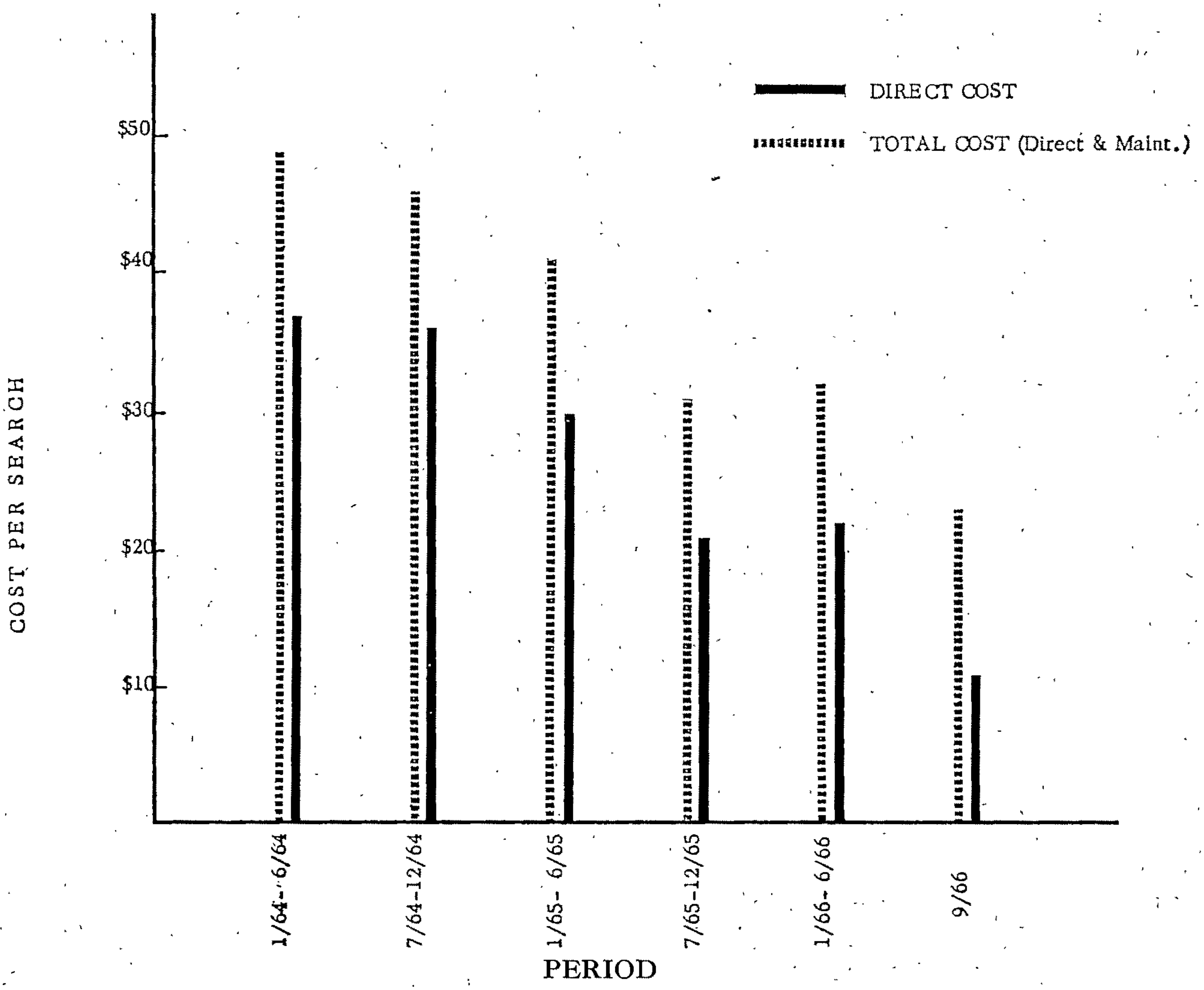


FIG. 2. Cost for computer searches

bibliographic searches will be accomplished via the disc files, (2) additional improvements are already being made to the operating systems, and (3) computer usage will increase thus decreasing the cost per hour and the prorated burden each task must carry for the maintenance expense.

Reference

1. MARRON, H., and M. SNYDERMAN, Cost Distribution and Analysis in Computer Storage and Retrieval, *American Documentation,* 17 (No. 2): 89–95 (1966).

Relevance Disagreements and Unclear Request Forms*

Disagreements about the relevance of documents to retrieval requests occur because relevance judges differently interpret requests or documents. Requests may be differently interpreted because they are unclear. Well-known types of request obscurity are reviewed. Less well known is that a request may be unclear because its form—"documents about subject *S*," "documents answering question Q," etc.—is unclear.

Explications are developed of the meanings of the request forms just given and several others. A request of any of the forms discussed is interpreted to be for documents which support statements of a specified kind in a specified way. For example, an "about *S*" request requires documents supporting statements which contain expression *S* (though several qualifications are needed); a "question Q" request requires documents which support answers to Q. Examples are given which suggest that some, perhaps all, "about *S*" requests are unclear. Some ways of formulating clear question requests are given.

Various ways in which documents may support statements are distinguished. These depend on such factors as parts of a document used, inference strength, and background knowledge permitted. Some possibly clear support specifications are indicated.

JOHN O'CONNOR †

Institute for Advancement of Medical Communication
Philadelphia, Pennsylvania

• 1. Background

Introduction

Document retrieval is the selection of documents of a specified kind from a large collection. A basically important way of specifying the kind of document desired is by subject. For example, "All theoretical papers on nuclear models," "What phosphorus compounds produce neurotoxic effects?" "Inconel and Monel: Composition, properties, weldability, and metallurgy of Inconel-*X*, Inco-*A*, Inco-140, and Monel," "What procedure can be used for preparing impurity-free ferric ethylate?" [1]

Subject document retrieval for scientists and engineers has received increasing attention in the last two decades. Traditional methods have been analysed further and many new methods have been introduced, some using computers. But it is not clear which methods are most effective, or whether any are effective enough. Therefore in recent years there has been a growing amount of work on testing retrieval methods and systems.

Relevance Disagreements

Documents which satisfy a subject retrieval request are usually called "relevant" to the request. Thus the basic function of a subject document retrieval system is to provide relevant documents for a request. However in practice there is sometimes disagreement among people competent to judge about whether a given document is relevant to a specified request. For instance, in a retrieval test, two different systems were used to retrieve from the same collection for the same ninety-eight requests. In comparing the retrieval results, the operators of the two systems agreed that 1,390 documents were relevant. But one group claimed relevance for an additional 488 documents and the other group claimed relevance for a different 1,089 documents. These claims were made after each group had examined all documents retrieved by each system (*4*). As another example, according to a survey of users of the American Society of Metals—Western Reserve University Metallurgical Searching Service, only about half the abstracts sent in response to

* This research was sponsored by the Information Systems Branch, Office of Naval Research (Contract Nonr N00014–66–C0096). Reproduction in whole or in part is permitted for any purpose of the United States Government.

† Present address: Center for the Information Sciences, Lehigh University.

[1] These examples are taken from four different lists of retrieval requests which are described in a table at the end of Part 2. The first example is from the AIP list and the others from (*1*, *2*, *3*), respectively.

requests were considered relevant by the users (*5*, p. 20) even though the output had been relevance checked by Service personnel.[2] "This is comparable with results of other systems" (*2*, p. 28). On the other hand, there was far more agreement among relevance judges in a retrieval experiment with physics literature (*6*). Fewer than 20 percent of the relevance judgments independently made by two physicists varied by more than two points on a scale of 10. There was then "resolution of all variances" through discussion with a third physicist. Moreover, "most of this variance was attributable to obvious oversight" (*6*, pp. 1100, 1104).

Relevance Disagreements and Unclear Requests

The results described in the first example of the preceding paragraph were interpreted by one of the groups involved, *ARC* (Armed Services Technical Information Agency Reference Center) in the following way:

> ARC took the position that, since it is difficult to be sure from a semantic viewpoint what a requestor wants, the Center prefers to send the maximum number of references to a requestor rather than take the chance of failing to include a reference which might be useful to him. On the other hand ARC pointed out that it appeared that Documentation Inc. acted upon the assumption that there exists for each question a relatively well defined area of pertinent information and that the boundary between a high plateau of pertinency and a surrounding lowland of irrelevant references can be located with considerable accuracy from the wording of the request (*4*, p. 236).

This fails to explain why ARC would not accept as relevant 488 documents judged relevant by Documentation Inc. However the passage is noteworthy for suggesting that relevance judges disagree because retrieval requests are unclear. A somewhat more explicit formulation of this idea is Mooers' assertion that, "Any inquiry from a customer can be interpreted in various ways. Depending upon the interpretation, the relevance of the documents produced will change" (*7*, p. 4).

Some evidence that relevance disagreements are caused by unclear retrieval requests is provided by the physics retrieval experiment (*6*) in which there was great agreement among relevance judges. A general assumption of that study was that

> the requester could in principle communicate his requirement with reasonable accuracy to some other person, especially to someone knowledgeable in the subject matter (*8*, pp. 281–282).

In particular, the questions devised for the experiment were

> highly specific and, to the extent possible, incorporate(d) within themselves the requester's "viewpoint" and "motive." Examples of such questions are: (i) What nuclear reactions are sensitive to the spin and parity of mesons and hence are useful in measuring those quantities? (ii) How does charge polarization within a nucleus affect the Coulomb scattering of charged particles by that nucleus? (iii) What are the "magic numbers" for nuclear shell structure? (*6*, p. 1100.) Should ambiguities of viewpoint arise it [was] postulated that the questions occur in the context of a college examination and that interpretation be based on best judgment applied under those conditions (*9*, p. 3).

Kinds of Unclear Requests

A number of different ways in which a retrieval request can be unclear are well known in documentation. They are summarily described below.

A request may contain an expression which is ambiguous in the situation. For example, a request "concerned insecticide in control of flies," and the requester judged some documents irrelevant because they were about black flies, etc., and by "flies" he meant houseflies (*10*, p. 138). As another example, the request, "Performance of engines with liquid injection," was searched by four people as part of a retrieval experiment. Three of the searchers understood "liquid injection" to mean "fuel injection," but the fourth interpreted it to mean "water injection," which was also the intent of the requestor (*11*, p. 180).

A request may be obscure because it is syntactically ambiguous. For example, would the request "Inconel and Monel: Composition, properties, weldability, and metallurgy of Inconel-*X*, Inco-*A*, Inco-140, and Monel" be satisfied by a paper on the weldability of Monel which said nothing otherwise about Monel and nothing about the other alloys named in the request? In other words, what do the *and*'s in the request mean? It seems to be usual practice to understand the *and*'s in a request like this as *and/or*'s. To the extent that is not commonly understood, this request and others of similar form are syntactically unclear.

A request may be unclear even though it contains no ambiguous subject-matter expressions and no syntactic ambiguity. For example, consider the request, "All theoretical papers on nuclear models." Would a paper in a journal of pure mathematics which solved a mathematical problem involved in a nuclear model without intending that physical application be relevant to the question? To take another example, for the request, "weldability of Monel," would a paper be relevant which described an efficient method of determining weldability for a class of alloys, including Monel, if the paper did not mention Monel? In general, is a request for papers about subject *S* satisfied by papers which, roughly speaking, only indirectly say something about *S*? If so, how indirect may the relation be? This kind of request obscurity might be called "vagueness of scope."

Aside from these problems in interpreting requests, there may be uncertainties related to what the requester knows and how he will evaluate papers. He may want only documents containing information new to him (perhaps including what he once knew but has forgotten), or he may want thorough retrieval for a comprehensive

[2] This figure includes an unknown percentage of abstracts judged by users to be relevant but not "useful" (*5*, p. A6).

review. He might wish only papers intelligible to him, or he may accept others as well because colleagues can help interpret them. Perhaps he desires only conclusive papers on the subject, or alternatively he is, for instance, a drug administrator concerned also with reports of possible side effects. He may wish only "significant" papers, or he may want everything he does not already know on the requested subject, depending perhaps on his general attitude toward use of the literature. Finally if, for instance, he asks for methods of preparing impurity-free ferric ethylate, he may be interested in any preparation methods, or he may want only methods using equipment he possesses; in general he may want any information on the requested subject or he may want only information useful in his particular circumstances.

Even if a requester indicates clearly how new, intelligible, conclusive, significant, and useful he wants information to be, it may still be uncertain what will be new, intelligible, or useful to him, because not enough is known of his background and circumstances. It may also be uncertain what he will judge conclusive, significant, or useful in cases where there can be competent scientific disagreement in such judging. For example:

> [At a meeting, following presentation of a paper] O. M. Reinmuth (University of Miami) raised questions concerning the validity of the scoring method, the experience of the individuals performing the various examinations, and the meaning of the increased jugular oxygen tensions which have been thought to show increased utilization of oxygen (*13*).

> . . . when, in 1879–1884, Georg Cantor communicated his fundamental results on [set] theory (now one of the bases of contemporary science), one of them looked so paradoxical and upset so radically all our fundamental notions that it unleashed the decided hostility of Kronecker, one of the leading mathematicians in that time, who prevented Cantor from getting any new appointment in German universities and even from having any memoir published in German periodicals. Of course the proof of that result is as clear and rigorous as any other proof in mathematics, leaving no possibility of not admitting it (*12*, p. 92n).

> [From an exchange of letters on an earlier paper] . . . their conclusion that acetazolamide is a potentially useful tool in teratology must be tempered by the following considerations . . . we find it difficult to temper our conclusion that acetazolamide is a potentially useful tool. . .(*14*).

Obscurities of the kinds described in the preceding two paragraphs will be called uncertainties of "user background." The name is only roughly appropriate for uncertainties concerning what a user will judge conclusive, significant, or useful, but that should cause no problem.

There is another way in which a retrieval request may be unclear. A request is sometimes interpreted as representing an interest only partly conveyed by the request's explicit formulation. For instance, a requester who asks, "How can Lissajous figures be generated by digital computer?" may be interested in a description of how they can be generated more exactly at comparable cost by analog computer (*15*). Or a user who asks, "What procedure can be used for preparing impurity-free ferric ethylate?" may be interested in a document naming a commercial chemical supplier of pure ferric ethylate.[3] In such cases, the document rather clearly does not satisfy the explicit request, though it may be of interest to the requester, thus the request obscurities of various types described earlier do not apply. However if an explicit request is "read between the lines" by several relevance judges other than the requester in an attempt to satisfy such an interest, they may disagree in interpretation of it. Or they may disagree in interpretation of the request because only one of them assigns an "implicit meaning" to it. Further, a requester may disagree with another relevance judge in regard to the "implicit meaning" of his request because he associates some implicit meaning with the request and the other judge does not, or vice versa, or the implicit meaning he associates with the request is different from that assigned by the other judge. Uncertainties of this kind in interpreting requests will be called obscurities of "implicit meaning."

There may be other general kinds of request obscurity besides those described above. An important one, obscurity of request form meaning, is the central concern of this paper (see Parts 2 and 3).

Other Possible Causes of Relevance Disagreements

Relevance judges may agree on the meaning of a request, but disagree about whether a particular document is of the kind specified by the request. There appear to be no instances of such disagreements described in the documentation literature. However, a relevance disagreement about a document might occur because the document is somewhat unclear, and is interpreted in different ways by different judges. Or there might be a relevance disagreement because the judges have different scientific intuitions about the paper. For instance, if a request is understood to ask for documents which are conclusive, significant, or useful for purpose *P*, rather than for documents which will be judged by the requester to be conclusive, significant, or useful for purpose *P* (as was assumed earlier in discussing user background uncertainties), then a disagreement about whether a particular document is, for example, conclusive is a disagreement about the document rather than about the request.

Some disagreements about relevance are the result of careless error in interpreting requests or documents. For example, in a case described previously, according to *ARC*'s analysis of its retrieval failures (492 papers missed by *ARC*, retrieved by Documentation Inc., and agreed by *ARC* to be relevant), twenty-five papers were missed "because the original interpretation of the request was inadequate" (*4*, p. 329). In the physics retrieval experiment described earlier, "most" of the initial relevance

[3] The phenomenon illustrated by these examples has sometimes been described by saying that requesters "do not know what they want" (*16*, pp. 15–16), or saying that a requester "may be unaware of his real need" (*17*, p. 81).

disagreements subsequently resolved by discussion were "attributable to obvious oversight" (*6*, p. 1104).

Lack of specialized knowledge on the part of one or more relevance judges, including the requester if he is searching outside his specialties, may lead to differing interpretations of requests or documents, and thus result in relevance disagreements. Several examples will be given later.[4]

It has been suggested, in this and preceding sections, that relevance disagreements may occur because judges interpret requests or documents differently, because of obscurity, disagreements in scientific intuition, carelessness, or ignorance. However, some studies of relevance disagreements have given quite different lists of possible causes. For example, Rees and Saracevic (*18*) suggest that relevance judgments are affected by such factors as the education and experience of the judge, his work functions (e.g., teaching, research, administration), the purpose he understands the request to have (e.g., solving a specific problem, compiling a bibliography), the environment (e.g., university, industry), the timing he understands the request to have (e.g., different stages in a research project), and the nature of the document representation (e.g., full paper, abstract). However, the two kinds of causes are related. For a judge may interpret a request or a document in a certain way because of such factors as his education, work functions, and environment, the document representation he is given, and what he understands to be the purpose, environment, and timing of the request. A similar remark applies to other lists of possible causes of relevance disagreements which have been given in the literature, insofar as such lists do not specify differences in request and document interpretation as possible causes.

The Basic Causes of Relevance Disagreements

The basic causes of relevance disagreements are differences in interpretation of requests or documents, rather than such factors as the education, etc., of the judges, and what they take to be the purpose, environment, and timing of the request. For if two judges agree on the kind of documents a request asks for, and agree on whether or not a particular document is of that kind, then their relevance judgments necessarily agree. On the other hand, if the judges are similar in education, etc., and agree in their understanding of the purpose, environment, and timing of the request, it is still at least abstractly possible for their relevance judgments to disagree.

Rees et al. (*18*) and Cuadra et al. (*19*) are empirically investigating how variations in relevance judges' education, etc., and understanding of a request's purpose, etc., are associated with disagreements in relevance judgments. Another worthwhile empirical investigation would be a study of whether particular relevance disagreements are caused by different interpretations of requests or different interpretations of documents, and whether the differences are caused by obscurity, disagreements in intuition, carelessness, ignorance, or other factors. How such a study might be conducted is sketched in the next paragraph.

Suppose relevance judges J_1 and J_2 disagree about the relevance of document D to request R. There should then be a discussion among the judges and a mediator. The judge who asserts relevance (say J_1) should be asked to paraphrase R (call the paraphrase R_1), and summarize the characteristics of D which match R_1 (call the summary S_1). He should be asked to formulate R_1 and S_1 so that the match between them is unquestionable; for instance R_1="any paper about A" and S_1="D is a paper about A." The judge who denies relevance (J_2) should then be asked what part of J_1's argument for the relevance of D to R he does not accept and why he does not. If he questions the paraphrase of R as R_1, then the original request R was unclear in some way, he or J_1 has made a careless or ignorant error, or they have interpreted R differently for some other reason. Further discussion should help to clarify which of these is the case. If J_2 accepts R_1 but questions S_1, then the disagreement is about the document. Further discussion should help to indicate its specific nature. The general structure of the "further discussions" which would follow initial disagreements about R_1 or S_1 need more study before this method of investigating relevance disagreements is tried. Consideration will also need to be given to minimizing construction of ex post facto arguments by judges to support judgments made on different grounds or no grounds.

• 2. Request Form Meanings

Unclear Request Forms

In Part 1 it was suggested that unclear retrieval requests may cause relevance disagreements. This raises the question of how retrieval requests can be expressed clearly. It might seem that a request with none of the obscurities described in Part 1 would be clear. However, this is not so. For example, even if the request, "papers on nuclear models," has none of those obscurities, relevance judges may still disagree about what constitutes being "a paper on" nuclear models. Similarly, even if the request, "What phosphorus compounds produce neurotoxic effects?" is otherwise clear, relevance judges may disagree about what constitutes being a document satisfying a question request. In general, the meanings of request forms appear to need clarification. Parts 2 and 3 of this paper describe the results so far of a study attempting such clarification. There seem to be no previous documentation studies of this problem.

Requests for Documents About a Subject

A traditional form of request is for documents "about," or "on" a specified subject, for example, "All theoretical

[4] Part 8, *Background Knowledge*, second paragraph.

question if it is a permitted substitution-instance of the question's statement-form.[11]

A general restriction on the substitutions in statement-forms of questions is that the results must be in well-formed English (or other discourse language). Therefore care must be taken in writing a statement-form that its structure not exclude as ill-formed some substitutions permitted by the original question. For example, if "What phosphorus compounds produce neurotoxic effects?" will accept as answers not only chemical names but also long descriptions of natural products, then "*X* is a phosphorus compound which produces neurotoxic effects" would not be a correct statement-form for the question. In general, when a question permits long descriptions to replace a variable, it may be adequate to formulate the question's statement-form with the variable at the end as a blank, allowing substitution of indefinitely long passages. As an example, the question above can be represented by the statement-form, "The following phosphorus compound produces neurotoxic effects: . . . ."

A question which seems clear may not be completely so, in the sense that it is uncertain which statements are answers to it. For instance, is the following statement an answer to the question, "What is the fastest add-time on current computers?"—"The fastest add-time on current computers is less than a microsecond and more than one-tenth of a microsecond"? As another example, is the following statement an answer to the question, "What adult monkeys are docile enough for laboratory use?"—"Adult stump-tailed macaque monkeys which are handled regularly are docile enough for laboratory use" (*22*)? If an attempt is made to represent either of these questions as a statement-form plus substitution conditions, the uncertainty of what constitutes an answer appears as an uncertainty about what the substitution restrictions are.

Suppose one tries to formulate completely clear questions by expressing them directly as statement-forms plus substitution conditions, rather than using natural language question formulations. The substitution conditions can be specified in various ways. One might permit any substitution-instance of the statement-form which is a well-formed statement of English (or other discourse language). A question so formulated is clear to the extent that "well-formed statement of English," for instance, is clear. More restrictive substitution conditions can be clearly formulated by explicitly listing which expressions may be substituted in the statement-form, or by naming an existing list of them (for instance, any drug listed in a particular pharmaceutical handbook). A different kind of substitution condition specifies the type of expression which may be substituted, without giving or naming an existing list. Some examples are: any natural number name, any chemical name, any monkey species name, any natural number name followed by "nanosecond" perhaps followed by "plus-or-minus" and the name of a natural number no more than 100. Some further examples are: any substance description, any description of a chemical laboratory procedure, any monkey species name accompanied by a description of a laboratory treatment of monkeys.

The examples in the last sentence of the preceding paragraph indicate that a substitution condition which specifies a permitted type of substituted expression may not be completely clear. For example, there are chemical procedures, close in magnitude to industrial processes, which competent judges might disagree about calling "laboratory procedures"; "laboratory treatment of monkeys" is similarly vague. As an example of a different kind, a paper reporting discovery of a new nerve fiber, say the Volk-Smyth fiber, would support the statement, "A phosphorus compound which destroys the Volk-Smyth fiber is a phosphorus compound which produces neurotoxic effects." Similarly, a paper reporting a new animal tranquilizer, say Calmatine, would support the statement, "Adult monkeys treated with Calmatine are docile enough for laboratory use." But competent judges might disagree about whether these statements are answers which satisfy the respective substitution conditions, "any substance description" and "any description of a laboratory treatment of monkeys." In general, techniques for clearly formulating substitution conditions need further investigation.

Some Extensions of Question and About Requests

A retrieval request which is simply a subject expression, for example, "Bent crystal spectrometers" (*23*, p. 47), is often interpreted as a request for papers about that subject, for instance, papers which say something about bent crystal spectrometers. However, in some cases a subject expression used as a request may be intended as a question request. For instance, "Precise measurements of *Q*-value via mass spectrometry and nuclear reactions both" (*23*, p. 47) may be a request for papers answering the question, "What values have been produced by precise measurements of *Q*-value via mass spectrometry and nuclear reactions both," rather than a request for papers saying anything about such measurements, for example, describing techniques for performing them. In general, requests which are subject expressions might be equivalent to either "about" or question requests.

Some requests have imperative form, for example, "List the types of pi mesons and explain why each must exist,"[12] and "Provide bibliography on thin films" (*3*, p. 51). The first of these examples appears to be equivalent to a question, "What are the types of pi mesons and why must each exist?" A number of imperative requests appear to be similarly equivalent to questions. The second example is a request for documents about thin films. "About" requests seem often to be imperatives.

[11] This explication of questions is suggested in (*21*, pp. 332–334). However, the authors treat yes-no questions by using a variable restricted to "true" and "false," rather than in the way illustrated by the bevatron example.

[12] An example from the data on which (*6*) is based. See the table at the end of the next section.

Some requests can be interpreted as combinations of simpler requests, for instance, "Which species of Aphididae attack leguminous crops in the United Kingdom and how can they be controlled?" (*24*, p. 30), or the "Inconel and Monel" illustration used earlier.

A request may give instructions for forming many different subject expressions, and can be understood as the disjunction of those expressions. A lengthy example (*2*, p. 49) is the following, which includes interchanges between the searchers and the requester:

> Metal—gas reactions at elevated temperatures.
> *Send:* the area under consideration here covers the broad area of surface reactions of metals with gases at elevated temperatures. This would include oxidation, pitting, scaling, thin film formation, tarnishing, etc. The metals and alloys considered would not be restricted to materials now used in the glass industry such as cast iron, platinum, etc., but would be all inclusive. Both the kinetics and mechanism of these reactions are of interest. With respect to specifying the gases to be considered, the search should be restricted to oxygen, air, nitrogen, water vapor, sulfur and sulfur compounds (inorganic, as sulfur dioxide and hydrogen sulfide), and mixtures of the above. Also of interest would be chemisorption of these gases on metals.

Presumably "kinetics of aluminum-oxygen scaling at elevated temperatures" is a subject expression specified by this request, and so is any expression obtained from it by making one of the indicated substitutions. The "etc." in one of the sentences of the request perhaps allows for an indefinitely long disjunction of subject expressions, and may help make the request somewhat unclear.

Some Data of Request Form Frequencies

Twelve available lists of retrieval requests were examined to find instances of request forms not yet considered. The instances found will not be discussed in the present paper. However, the frequencies of various request forms in the lists are given in the table at the end of this section. Several things should be said about how the table was compiled. A request was classified "Other" if it contained, roughly speaking, "documental" terms rather than just subject-matter terms. Some examples are: "What research and development is going on in the field of engines and automotive components involving new lubricants, fuels, and power transmission fluids?" (*3*, p. 56), in which "in the field of" is a documental expression, and "List all papers involving the study of human volunteers" (*1*, p. 64), in which "papers involving the study of" is documental. Exceptions to this procedure were requests explicitly asking for documents on or about a subject, such as "Prepare bibliography on the stress corrosion of steel" (*3*, p. 58), or "Is there a recent book or paper, preferably in English, on the development of rockets?" (*24*, p. 32). Nine imperatives which seemed equivalent to questions were counted as questions. Compound requests were counted as having the form of any constituent request; there were no mixed compounds. Requests which gave instructions for forming disjunctions of subject expressions (all in 16) were classified as subject expression requests. A request which consisted of just a subject expression was so classified, without an attempt to interpret it as equivalent to either a question or an "about" request (Table 1).

The requests in Table 2 are based on examination of every tenth request in each list. The number in parentheses in each case is the randomly selected number of the first request examined.

3. Document-Statement Inferences

Introduction

In Part 2 retrieval requests of several forms were interpreted as requests for documents supporting statements of specified kinds. However a document may support a statement in a variety of different ways. Some of these are described in this section.

For this discussion it will be assumed that a judge with appropriate competence is asked to decide whether a particular document D supports statement P by an inference of type I. An inference type I is specified to the judge by an instruction, for instance, "Accept the paper uncritically, use any physics as background knowledge, and infer P conclusively." The example just given is a composite inference type, while the kinds of inference

Table 1

Source	Request Form				Circumstances of request
	Question	Documents about a subject	Subject expression	Other	
2, pp. 47–66	—	—	16	3	real, to operational system
1, pp. 64–65	22	—	1	2	?
23, pp, 47–49	—	2	33	15	real, to hypothesized ideal system
24, pp. 30–32	17	6	—	9	some real
6[13]	48	—	—	2	invented
25, pp. X6–X8	6	—	10	1	invented

[13] For access to the data on which (*6*) is based, the author is grateful to Christine Montgomery and J. L. Kuhns of the Bunker-Ramo Corporation.

TABLE 2

Source	Request Form				Circumstances of request
	Question	Documents about a subject	Subject expression	Other	
3, pp. 43–62(8)	20	4	—	7	real, to hypothesized ideal system
AIP[14]	—	2	13	12	real, to hypothesized ideal system
11, pp. 130–134(7)	—	—	10	—	invented
26, pp. 84–91(4)	14	—	—	—	invented
27, pp. 185–200(2)	17	2	1	8	real
28, pp. 781–796(4)	4	—	4	2	invented

[14] For access to all the retrieval requests of which a sample is given in (*23*, pp. 47–49), the author is grateful to Pauline Atherton of the American Institute of Physics.

described below will be elementary (at least apparently so). Each elementary type of inference will be described in the form of an instruction.

Author Intent

The support of a particular statement by a document need not have been intended by the author. Authors sometimes overlook even important and obvious consequences of their work. For example:

> Two theorems, important to the subject, were such obvious and immediate consequences of the ideas contained [in Hadamard's thesis] that, years later, other authors imputed them to me, and I was obliged to confess that, evident as they were, I had not perceived them (*12*, p. 51).
>
> [A chemist] had done some experimental work in 1955 and had published a report without fully realizing the relevance of his work to the chemical theory of a certain reaction mechanism. Between 1955 and 1957, he was led to earlier literature which suggested this significance of his work to him. During the same period, this fact was also brought home to him through three contacts with other scientists which had ensued from his work in three quite independent ways (*29*, p. 207).

The inference judge can be instructed to infer statement *P* from document *D* only if *P* was intended as a consequence of *D* by *D*'s author. However, in some cases this may be difficult or even impossible to decide. The Hadamard example illustrates that competent readers of a document can sometimes be mistaken about what its author intends. Therefore, alternatively, the inference judge might not be given any instruction concerning author intent.

Inference Basis

The judge may be instructed to accept all of document *D* as a basis for inference, or to accept only certain parts of it (for instance sections headed "Methods" and "Results"), or certain kinds of statements in it (for example, descriptions of what was done and observed). If kinds of acceptable statements are specified, the specifications may not be completely clear. For example, competent judges may disagree about "what was done and observed" in a particular experiment, even if it is reported in a well-written paper. An illustration of this is the following:

> Consider two microbiologists. They look at a prepared slide; when asked what they see they may give different answers. One sees in the cell before him a cluster of foreign matter; it is an artifact, a coagulum resulting from inadequate staining techniques. This clot has no more to do with the cell, *in vivo*, than the scars left on it by the archaeologist's spade have to do with the original shape of some Grecian urn. The other biologist identifies the clot as a cell organ, a "Golgi body." As for techniques, he argues: "The standard way of detecting a cell organ is by fixing and staining. Why single out this one technique as producing artifacts, while others disclose genuine organs." [15]

The judge may also be instructed, concerning any portions of *D* he is not asked to accept immediately, to admit them to the inference basis if he thinks they are reliable. He may further be asked to add qualifying phrases where he thinks such addition will make passages reliable. Examples of such phrases are "perhaps," "for the user population sampled," "in that sense of 'Algol-like,'" "if one can assume consistency," etc.[16] Such modified passages are then also to be added to the inference basis. Note that two competent judges may disagree about what passages in a document are reliable, and what critical annotations are necessary. This is illustrated by the examples given in Part 1 of disagreements about the conclusiveness, significance, and usefulness of documents.[17]

For the rest of Part 3, any reference to inferences from document *D* shall mean inferences from a basis derived from *D*.

Type and Strength of Inference

The judge may be instructed to determine only whether *D* formally implies *P*, or to consider also the possibilities

[15] Hanson cites "the papers by Baker and Gatonby in *Nature*, 1949 to present" (*30*, p. 4).

[16] This kind of annotation only decreases (roughly speaking) what can be inferred by *D*. Annotations which increase *D*'s power are background knowledge augmentations, which will be discussed later.

[17] Part I, *Kinds of Unclear Requests*, sixth paragraph.

that P and D are connected by a mathematical statistical argument, or by a nonmathematical probable inference. The last mentioned type of reasoning is illustrated by the inference from biochemical information and animal test results that a particular drug is safe for human testing, or the inference that a physical constant has a certain value because measurements of it by several independent methods have given closely agreeing results.

If the judge is instructed to look for a probable inference of either kind, he must also be told how strong it must be. For nonstatistical probable arguments there appears to be no precise language for specifying inference strength (see, for instance, *31*, p. 212). Only an approximate language seems to be available, using such expressions as "a bit of evidence for," "moderately supports," "makes highly probable," etc. It is uncertain how clear this language is in various circumstances. It should also be noted, concerning informal probable inferences, that two competent judges may disagree about the strength of a particular inference. Examples of such disagreement were given earlier.[18] An illustration of a more general kind is the following passage quoted by Polyanyi (*32*) from the directions to Royal Society referees:

> A paper should not be recommended for rejection merely because the referee disagrees with the opinions or conclusions it contains unless fallacious reasoning or experimental error is unmistakably evident.

The judge may be instructed to determine, supposing D does not formally imply P, whether D formally implies a probabilistic assertion of P, such as "It is fairly probable that P" or "It is a plausible conjecture that P." Such a statement might be implied if, for instance, the original document D only advanced P as a conjecture, or if an unqualified assertion of P by D has been weakened by a critical annotation of the inference judge. If the judge is to look for such an implication, he must be told how strong the probability attributed to P must be. For nonstatistical situations this specification can apparently only be approximate and perhaps somewhat unclear, as for the nonmathematical probable inferences referred to in the preceding paragraph.

In summary, the judge should be told what kinds of inference to consider (formal logical, statistical, informal probable), and how strongly P must be supported by whatever kinds of inference may be used.

Background Knowledge

The inference judge might be instructed to use no background knowledge in attempting to infer statement P from document D. However to draw any inference at all from a document in natural language, he would have to use his general knowledge of the language. For instance, he could not otherwise infer "Lethenone is neurotoxic" ("Lethenone" is a fictitious insecticide name) from such passages as "This neurotoxicity of Lethenone implies that it should not . . . ," or "Is Lethenone neurotoxic to humans? The answer is definitely in the affirmative." Moreover, if he were instructed to use no knowledge of language but only to look for the word sequence, "Lethenone is neurotoxic," as a guide to inferring that statement, then he would incorrectly infer it from a document in which, for instance, the word sequence, "Lethenone is neurotoxic," appeared in a series of clauses preceded by the expression. "The following false statements have appeared in the press:" No attempt will be made here to define the notion "general knowledge of the language." Note, however, that such knowledge is not only syntactic (unless "syntactic" is defined extremely broadly), since it also includes knowing the specific meanings of such words as "answer," "affirmative," and "false." How "general knowledge of the language" should be defined for purposes of request form clarity needs further study.

To infer a statement from a document may require knowing meanings of particular subject-matter expressions. The expressions might be verbal, for instance, "neurotoxic" or "nuclear reaction parameters," or they might be nonverbal, for instance, equations or diagrams. Subject-matter knowledge which is not merely of expression meanings may also be needed to draw an inference. For example:

> A physiologist should know that a hypothermic dog has a slow heart rate and should index Bradycardia even if the author discusses Heart Rate. A chemist could justifiably be unaware of that fact (*33*, p. 29).

> The physicist's function as an analyzer of the [retrieval request] data has been found to be largely in two areas:
>
> First, to improve the precision of those search requests where improvement was needed, e.g., "advances in the development of the optical model (experimental and theoretical)"—since the optical model is a *theoretical* method for calculating cross-sections, angular distributions and polarizations, the *experimental* portion of this request would necessarily relate to determinations of these parameters for nuclear and incident particle energies for which optical model calculations are valid (*23*, p. 18).

Suppose document D does not support statement P by a formal implication or a mathematical statistical argument, and the inference judge is supposed in such a case to determine how much D supports P by nonmathematical probable inference. He cannot determine this in a vacuum, in the way that a formal logical or mathematical statistical argument can be developed and examined. Instead he must have and use a considerable amount of background knowledge in the subject field(s) of document D and statement P. For if he infers P from D with some probability, he must know enough to be assured that there is no sufficiently plausible way D can be true and P false. On the other hand, if he judges D to support P with less than a certain strength, he must be assured that he has not overlooked some likely enough connection between D and P which someone knowledgeable in the subject could point out. Thus if the inference judge is

[18] Part 1, *Kinds of Unclear Requests*, sixth paragraph.

not allowed or not able to use a full range of subject knowledge (of a not easily specified extent) as background knowledge to augment document *D*, he is restricted to formal logic and mathematical statistics in attempting to infer *P* from *D*, even from *D* augmented with whatever background knowledge he is able to use. For similar reasons, if the inference judge cannot use a full range of subject knowledge (of a not easily specified extent), his role in deriving an inference basis from *D* is limited. He cannot judge for himself the reliability of passages in *D*, and he cannot add critical annotations to *D*, except those concerning logical or statistical arguments within *D*.

Suppose some subject-matter knowledge may be used by the inference judge. How clearly can the knowledge he may use be specified to him? A list of books would omit current knowledge (including that which corrects errors in books), and adding a list of journals would still omit "invisible-college" current knowledge. But the requestor may want current knowledge from some fields to be used. How clearly it can be specified needs further study. A different way of specifying what subject knowledge, especially knowledge of expression meanings, may be used by the inference judge is to provide a "thesaurus" which lists technical words and phrases, and gives other technical expressions which may replace them. However, a conventional thesaurus may not be immediately usable by an inference judge. For example, a typical thesaurus entry is *"bananas:* use *food,"* but "Monkeys do not require food to live" hardly follows from "Monkeys do not require bananas to live." Similarly, *"Pennsylvania:* also coded *United States"* does not warrant inferring from "Harrisburg is the capital of Pennsylvania" that Harrisburg is the capital of the United States. In general, the subject knowledge which the inference judge may use might be specified to him in a thesauric or other codified form, rather than in natural language texts. But investigation is needed of how this can be done clearly and correctly.

Some Possibly Clear Support Specifications

The instructions to the inference judge might be clear if they specify, for example, the following:

Pay no attention to author intent
Use the whole document as an inference basis (or sections of it with certain headings such as "Methods" and "Results")
Consider only formal implication of *P* by *D* (or mathematical statistical inference deriving *P* from *D* with at least a specified probability)
Use as background knowledge only "general knowledge of the language"

In the preceding sentence, "might" can be changed to "will" if a definition of "general knowledge of the language" can be formulated which is clear enough to prevent disagreements among inference judges. This needs further study.

• 4. Some Remarks

A subject document retrieval system which provides documents supporting specified kinds of statements in specified ways is not a "statement retrieval" or "fact retrieval" system. For a statement retrieval system provides statements of specified kinds which are supported in some way (often system-specified) by the corpus of the system. A document retrieval system, on the other hand, provides documents from which the requestor or some other reader external to the system must infer the statements of specified kinds. Thus, a document retrieval system does less than a statement retrieval system. On the other hand, suppose a document retrieval system successfully permits a requestor more freedom in specifying the kind of support for statements which he wants than does a statement retrieval system. Then in this dimension the statement retrieval system does less than the document retrieval system.

A basic emphasis of this paper has been that clear retrieval requests are important. But it is often asserted that a searcher is unavoidably "unclear" about what he wants, especially if he is searching in a somewhat unfamiliar field, as is often the case. Therefore, it is also asserted, he needs the assistance of a retrieval system expert, a cross-reference system, search cycling, or all three. However, there need be no inconsistency between these two viewpoints. If it is generally possible to formulate clear requests, this should help rather than hinder dealing with an initial uncertainty about what the requestor wants. In this context it should be noted that "unclear about what he wants" is ambiguous between, for instance, not formulating a clear question request and not knowing what answers will be supported by documents satisfying a clear question request. This ambiguity seems often to be overlooked in documentation literature.

Several remarks might be made about the kinds of unclear requests described in Part 1. Subject retrieval systems usually discourage clear language, for request form meanings are unclear; however, if request forms have clear meanings, then ambiguous subject expressions and syntactic ambiguity may be no more frequent in retrieval requests than in other kinds of scientific communication. "Vagueness of scope" in Part 1 referred to, roughly speaking, how indirectly a document may say something about a requested subject; a request does not have such an obscurity if it specifies what kinds of document-statement inference, especially what kinds of background knowledge, may be used. How "user background" requirements in a request can be formulated clearly needs very much study; the concepts of statement kind and document-statement inference kind may be more helpful in such a study than the usual documentation language of subjects; for example, a requestor might be better able to indicate what will not be new to him in terms of kinds of statements than in terms of subjects. "Reading between the lines" of a request seemingly unavoidably leads to differing interpretations of the request; to prevent confusion, documents retrieved to satisfy a clearly formulated request should be kept separate from additional documents volunteered as a result of "reading between the lines" of the request.

In the physics retrieval experiment of Swanson et al. (6), there was little disagreement among relevance judges, and a special attempt was made to formulate requests clearly.[19] However, the requests were not accompanied by inference instructions, the question requests (as most of them were) were not accompanied by substitution condition statements, and there was only a sketchy indication of "user background" requirements ("the context of a college examination"). Presumably in this experiment there was enough unstated intellectual accord among the relevance judges that such explicit formulations were unnecessary. It is not clear whether this was because they were all nuclear physicists, or for more specific reasons. In any case, such prior intellectual rapport cannot always be safely assumed. When it cannot, more explicit formulations of requests are needed.

The discussion in Parts 2 and 3 centered on statement types and document-statement inference types. It is suggested that, for a number of documentation problems, these concepts may be clearer and more helpful than the usual documentation notions such as subjects, relations between subjects, pertinence of subjects to documents, etc. For instance, they may be more useful in formulating rules to guide subject indexers. The distinctions among various kinds of document-statement support may help clarify the dispute about how much subject knowledge indexers must have. To the extent that the ideas in Parts 2 and 3 will permit clear subject requests to be formulated, the testing of retrieval systems might be less complicated by relevance disagreements. And the statement and inference concepts may be more helpful than traditional documentation notions in specifying what a computerized retrieval system is to do. If any or all of these suggestions turn out to be true, it should not be too surprising. For in the sciences it appears that a document is usually important to a requestor because it makes certain statements and provides a basis for inferring certain other statements. He cannot use subjects from the literature in his work, but he can use statements.

References

1. Goldwyn, A. J., Searching the Medical Literature, *Methods of Information in Medicine*, 2(2): 59–65 (1963).
2. National Academy of Sciences, Ad Hoc Committee of the Office of Documentation, *The Metallurgical Searching Service of the American Society for Metals—Western Reserve University: An Evaluation*, Publication 1148, National Academy of Sciences, Washington, D. C., 1964.
3. Smith, C., *User Requirements for Chemical Information and Data System*. Report R–1755, Frankford Arsenal, Philadelphia (April 1965).
4. Gull, C. D., Seven Years of Work on the Organization of Materials in the Special Library, *American Documentation*, 7 (No. 4): 320–329 (1956).
5. Wayne, I. *A Survey of Users of the American Society of Metals*, Western Reserve University Searching Service, Bureau of Social Science Research, Washington, D. C., 1962.
6. Swanson, D., Searching Natural Language Text by Computer, *Science*, 132 (34): 1099–1104 (October 1960).
7. Mooers, C., The Intensive Sample Test, Zator Co., Cambridge, Mass. (August 1959).
8. Swanson, D., Research Procedures for Automatic Indexing, In *Machine Indexing*, American University, Washington, D. C., 1962, pp. 281–304.
9. Swanson, D., *An Experiment in Automatic Text Searching*, Rep. No. C82-OU4, Ramo-Wooldridge Corp., Canoga Park, Calif. (April 1960).
10. Ferguson, M., The Communicable Disease Literature Project, In *Information Retrieval in Action*, Press of Western Reserve University, Cleveland, Ohio, 1963, pp. 135–140.
11. Cleverdon, C., *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*, College of Aeronautics, Cranfield, England (October 1962).
12. Hadamard, J., *Psychology of Invention in the Mathematical Field*. Princeton University Press, 1945, reprinted by Dover, New York, 1954.
13. Milliken, C. H., Cerebrovascular Disease, *Science*, 152: 805 (May 6, 1966).
14. Maren, T. H., and W. M. Layton, Letters, *Science*, 150: 79 (October 1, 1965).
15. Potzick, J., Lissajous Figures by Analog Computer, *Science*, 149: 1446 (September 24, 1965).
16. Rees, A., Remarks, in *Faceted Classification Schemes*, by B. C. Vickery, Graduate School of Library Service, Rutgers University, New Brunswick, N. J., 1966, pp. 15–16.
17. Bryant, E., Schema of Sources of Failure in IR Systems and Their Consequences, In *Study Conference on Evaluation of Document Searching Systems and Procedures: Summary*, National Science Foundation, Washington, D. C. (PB 166–905), B1–2 (February 10, 1965).
18. Rees, A., and T. Saracevic, *The Measurability of Relevance*, paper presented at the October 1966 meeting of the American Documentation Institute, Santa Monica, Calif.
19. Cuadra, C., et al., *Experimental Studies of Relevance Judgments: Second Progress Report*, Rep. No. TM-2068/000/00, System Development Corp., Santa Monica, Calif. (July 1966).
20. Goodman, N., About, *Mind*, 70 (277): 1–24 (January 1961).
21. Lewis, C. I., and C. H. Langford, *Symbolic Logic*, Century, New York, 1932, reprinted by Dover, New York, 1959.
22. Orbach, J., and A. Kling, Letter, *Science*, 148: 1173–1174 (May 28, 1965).

[19] See Part 1, *Relevance Disagreements*, and *Relevance Disagreements and Unclear Requests*.

23. ATHERTON, P., *A Preliminary Report on Phase 1 of the Reference Retrieval System Development Project,* Report No. AIP/DRP 62-2, American Institute of Physics, New York, (April 1962).

24. WOOD, D., and K. BARR, Courses on the Structure and Use of Scientific Literature. *Journal of Documentation,* 22(1): 22–32. (March 1966).

25. ROUCHIO, J., and M. ENGEL, Test Design and Detailed Retrieval Results, in SALTON, G., ET AL., *Information Storage and Retrieval.* Rep. No. ISR-8, Computation Labs, Harvard University, Cambridge, Mass., pp. X63X8 (December 1964).

26. CLEVERDON, C., and J. AITCHISON, *A Report on a Test of the Index of Metallurgical Literature of Western Reserve University,* College of Aeronautics, Cranfield, England (October 1963).

27. CLEVERDON, C. J. MILLS, and M. KEEN, *Factors Determining the Performance of Indexing Systems,* College of Aeronautics, Cranfield, England, 1966.

28. WRIGHT, R., and C. WILSON, Classification with Peek-a-boo for Indexing Documents on Aerodynamics: An Experiment in Retrieval, in *Proceedings of the International Conference on Scientific Information,* National Academy of Sciences, Washington, D. C., Vol. 1, 1959, pp. 771–801.

29. MENZEL, H., Planned and Unplanned Scientific Communication, in *Proceedings of the International Conference on Scientific Information,* National Academy of Sciences, Washington, D. C., Vol. 1, 1959, pp. 199-245.

30. HANSON, N. R., *Patterns of Discovery,* Cambridge University Press, Cambridge, England, 1958.

31. TOULMIN, S., *Foresight and Understanding,* Indiana University Press, Bloomington, 1961.

32. POLANYI, M., *Personal Knowledge,* University of Chicago Press, 1958, p. 145n.

33. WELT, I. D., and J. T. MACMILLAN, A Study of Indexing Procedures in a Limited Area of the Medical Sciences, *American Documentation,* 12 (No. 1): 27–31 (1961).

A Study of the Use of Materials Circulated from an Engineering Library, March to May 1956

The purpose of this study undertaken from March to May 1956 was to determine how an academic engineering library was used by two groups of users, the undergraduates and graduate students–faculty. The part of the study reported here is the result of a questionnaire given to the user at the time he charged out an item at the circulation desk to ascertain for what purpose he selected the item(s) and how he learned about it as a source of information. Variations in the reasons for selecting items changed during the three periods the data were collected. The undergraduates borrowed less for classroom work as the semester progressed, while the graduate-faculty group borrowed increasingly for this reason. Supporting the conclusions of the other studies, the most important source for learning about an item was personal through recommendation; however, one of every four items charged out was discovered in browsing through the library's collection. From this study one can conclude that not only must librarians be acquainted with their users as individuals, but that the physical arrangement of library materials is an important factor in accessibility to information in an academic environment.

VERN PINGS †

University of Wisconsin
Madison, Wisconsin

The increase in the amount and variety of published materials during the last decade has been the subject of considerable discussion by college and university librarians. An equally dramatic, but perhaps less discussed and appreciated problem is the change that colleges and universities have undergone since the war and the changes which will have to take place if those educational institutions are going to cope with the growth of our population and our material culture. The curricula of our schools have to be planned to educate students to a technological world that was hardly believed possible even fifteen years ago. Chemicals that were only a test tube curiosity a short time ago are now part of our daily lives. The majority of medicines now sold over the drug counters were not listed in the official Pharmacopoeias ten years ago. The science of electronics has changed the social pattern of our homes through television and other devices; our industrial production has already become completely automatic in a few industries. Not only have the schools been required to provide new laboratories, libraries, and physical plant, but the educators and administrators now must give instruction to increasing numbers of students doing graduate work. The College of Engineering at the University of Wisconsin, which is certainly not atypical, has had an increase of 600 percent in its graduate enrollment, while the undergraduates have almost doubled in the past ten years. Similarly the number of research projects presently being carried on is over 200 as compared to approximately 20 in 1939.

These changes in the curriculum, student body, and faculty have profoundly affected the organization and services provided by the Engineering Library. New emphases and new organization has become mandatory. Present space is inadequate. But because plans are now under consideration for the construction of a new library (still some four to six years hence), expensive renovation would not be accepted budget-wise. Any expenditures for capital equipment must be, or at least ought to be, made with the consideration that the equipment be usable in the new library. Since the new library is now in the beginning planning stages, questions arise as to how it should be constructed and organized to meet the future needs of the College of Engineering and other organizations and institutions it now serves. Because of the many unanswerable

† Present affiliation: Wayne State University, Detroit, Michigan.

questions, a two-year study was planned to determine if data could be obtained to solve some of these problems with assurance rather than just by empirical guesswork.

• Description of Project

In organizing the project, the functions of the library were divided into two aspects: (*a*) from the viewpoint of the librarians, and (*b*) from the viewpoint of the users. Or, to phrase it in another way, the functions were divided to determine the efficiency of the techniques and routines and the more intangible aspects of determining how the library was understood and used by the patrons.

For the first part of the study, arrangements were made with the instructors of the Time and Motion Study classes to use the library as a laboratory. At the time of this writing the study of the faculty and students has pointed up many practices which can be carried on more efficiently. However, it is still too early in the project to obtain information on general methodolgy and techniques.

An examination of the Engineering Library showed that the way in which the collection was used could be divided into three categories: (*a*) the reserve and reference collection, (*b*) the materials used within the library other than those items on reserve or reference, and (*c*) the materials checked out for use outside the library. It is with the latter category this paper is concerned.

This aspect was chosen first for study for several reasons. How the patrons become acquainted with and for what purpose they use the reserve collection are fairly well known. The more important question was how to provide improved service for the reserve collection, yet integrate the actual work with other aspects of service that need to be improved or expanded. The physical arrangement of the library and the shortage of personnel precluded any attempt to survey the use made of materials within the library at this time. Second, it was felt that studying this aspect first would give a good view of the patron approach to the library. Herner's study on the "Information Gathering Habits of Workers in Pure and Applied Science" (*1*) showed that of the 600 scientists interviewed at Johns Hopkins, only 11 percent did most of their reading in libraries even though 42 percent of this group depended primarily on the library for published materials and 49 percent depended equally on the library and their personal collections. Herner also points out that engineers did the least amount of reading in the library. Bernal (*2*) found that 56 percent of the scientists responding to a questionnaire obtain their materials from a library and that one-third of the papers studied were taken out of the library. Although the findings of Herner and Bernal, cannot be applied directly to the situation under observation here, the statistics they give do indicate that data obtained from individuals checking out books and other materials from the library would reveal to a considerable extent the way the library collection is used.

• Method of Study

In selecting a method for gathering data two possibilities were considered: interviews and questionnaires. Although interviews are decidedly the better method for obtaining information for this type of survey, insufficient staff prevented us from adopting this method exclusively. If mailed questionnaires are used, there is always the danger of an unrepresentative sample. The method finally selected was comparatively simple and direct. A questionnaire was prepared (see Table 1) which was given to the patron at the time he was charging out an item. The patron then could ask questions if there was doubt in his mind about how to fill it out, and the patron could be interviewed on his responses. This method had the advantages of a questionnaire where answers are consistent; the ambiguities could be clarified through an interview, and the person filling out the questionnaire did not have to rely on his memory in giving his answers—

TABLE 1. Data obtained and questions asked

You have selected this item you are now charging out

______I. for classroom work

______II. because it is applicable to your research project or thesis

______III. to provide you with information in or allied to your major field

______IV. because it is of interest to you for reasons other than those above

You became acquainted with this item

through discussions

______V. with members of your department or project

______VI. with professional contacts other than a member of your department or project

______VII. with someone other than a person connected with your work

______VIII. through an advertisement, review, or book announcement

______IX. through browsing in the library collection

______X. through a bibliography

______X. you found in a book

______XI. you found in a magazine

______XII. especially prepared for a class or a subject field

______XIII. through a subject heading in the Engineering Library Card Catalog

______XIV. through some regularly published index or abstracting publication, e.g., the *Engineering Index*, *Chemical Abstracts*, etc.

he was questioned on his purposes and methods at the time he was involved in selecting an item. This avoided the necessity of asking an interviewee to "guess" in his answers. In spite of this apparent convenient arrangement, more than a few questionnaires had to be discarded because patrons disappeared before ambiguous answers could be clarified through an interview.

Because the patron may charge out materials without the assistance of library personnel at the Engineering Library, not every individual who charged out material was asked to complete the questionnaire. Consequently considerable care was taken to acquire a "random" selection of individuals. During the ten-week period of the survey (March 6 to May 15, 1956) every effort was made to get questionnaires throughout the time the library was open. Only two individuals refused to fill out the questionnaire. The reason they gave in each instance was that they were due at a class.

The questionnaire was given only to those individuals connected with the University of Wisconsin. Materials charged out to individuals, organizations, or institutions other than those affiliated with the University constitute less than 5 percent of the total non-reserve book circulation. Further it was felt that materials circulated outside the University could be ignored because the primary concern was to determine how to improve the Engineering Library to meet the needs of the student body and faculty.

• Factors Affecting the Use of the Engineering Library

To get a picture of the results of this survey, the policies on the use of the library and its physical organization must be understood. As was already mentioned, the patron is permitted to charge out materials without the assistance of library personnel. All items in the Engineering Library may be borrowed for a period of one month (faculty members have a six months' borrowing privilege) except those items on the reserve and reference shelves and current periodicals. Even the latter may be kept for a month or longer if the patron requests a loan of that length of time and no other individual asks for the item. The library has open stacks and the only items which the patron cannot obtain himself are microfilms which are kept in a room with the microfilm reader and the librarian's reference works which are kept in the librarian's office. There is nothing sacrosanct about the librarian's office or the microfilm reading room and once the patron is aware of the physical location of these items, he may use them without any special permission from the library staff. Although these two categories of materials are the only ones not "readily available" to the library patron, the groupings and arrangement of other materials does have an effect on their use. Librarians often forget that the patron must know a great deal about the organization of a particular library before he can use it effectively—the librarian's view of the physical organization is that it is logical and sensible and consequently obvious. The organization of materials may not be obvious to the patron if he is expected to locate items for himself. Other than the two categories mentioned above, the patrons of the Engineering Library must be acquainted with the following groups:

1. *Periodicals and continuations.* Bound periodicals and continuations are arranged alphabetically by the most recent titles (Cutter numbers are assigned each file). Unbound periodicals are kept on a separate set of shelves until ready for binding. The continuations on the other hand are placed in pamphlet boxes beside the bound numbers. Not all publications which might be defined as a periodical or continuation are in this collection; some are classified and kept with the "book" collection.

2. *Books.* The books are divided into three groups: (*a*) the new and/or uncataloged books, (*b*) the book in the Cutter classification, and (*c*) those in the Library of Congress classification. The University of Wisconsin libraries changed to the Library of Congress classification system in 1953. From the librarian's viewpoint, the Public Catalog cards are clearly marked to show in which group a book may be found, but the distinction is far from obvious to many patrons.

3. *Reference and reserve materials.* These are kept in one area. Only the main entry card in the Public Catalog gives the location of these items.

4. *Indexes and abstracts.* All of the current indexing publications the library subscribes to are located in this special grouping. Only the publications devoted entirely, or almost entirely, to the publishing of abstracts are placed in this category; e.g., the abstracts published as a part of a professional journal are bound and kept with that periodical.

5. *Theses.* The theses are shelved separately and a separate catalog is maintained for them.

6. *Pamphlet collection.* Materials in this category are placed in pamphlet boxes. There are subject references of this collection in the Public Catalog.

• Definitions, Limitations, and Scoring of the Questionnaire

When a patron charged out more than one item at a time, the responses obtained from the questionnaire were treated as a unit, unless the patron indicated the items he was charging out were to be used for different purposes or he became acquainted with them through different means. We were interested in determining how the individual patron used the library rather than how separate items of the library were used or located.

Often the patron became acquainted with an item through a series of steps. For example, a patron, while scanning a current periodical, found a reference to a book in the bibliography of an article. He located the book in the library, but discovered it would not be of interest to him, but on browsing through the books classified in the same group, he found one that did interest him. In such an instance, the last step was the one used in scoring the questionnaire.

Because of difficulties in distinguishing clearly between faculty and graduate students, the responses of these two groups were combined. This difficulty arises from the fact that research assistants, fellows, etc., have the same privileges as faculty members in the use of the library. Oftentimes the graduate would not indicate his faculty status or a faculty member did not indicate that he was doing graduate study.

Although no one who filled out the questionnaire questioned the meaning of "research," the scoring of the questionnaire might be confusing without some further definition. A large number of undergraduates indicated that the reason for selecting an item was because it was applicable to their research or thesis. The only undergraduate group in the College of Engineering required to write a thesis are those in Civil Engineering. For the thesis to be accepted it is not necessary that "original" research be done as might be defined for a thesis to complete the requirements for an M.A. or Ph.D. degree. However, this ambiguity in the interpretation of the meaning of research is lessened if it is realized that the questionnaire attempted to obtain individual responses and that research, even in Engineering, does not necessarily always have to include experimentation. The undergraduate student who prepares a paper as a requirement for his class work may be doing research in the sense that he is relating ideas and concepts which are new to him, and for that matter may never have been related before. Because few papers prepared by undergraduates are published does not mean that the ideas in them are that much less valid or important than many of the papers that do get published. To the undergraduate, his paper is a means of classifying and organizing knowledge which is the major part of any research project. Since our purpose was to try to gain some insight into how the user viewed the functions of the library in relation to his own work, a more rigid definition of research did not seem necessary.

The questionnaires were grouped into three time periods, the first of four weeks and the last two periods of three weeks. The four-week period included the spring recess.

• Analysis of Data

The only other survey known to have been published that might be compared to this one is that of Urquhart at the Science Museum Library in England (*3*). A questionnaire was sent only to those individuals borrowing through the mails, that is, on an interlibrary loan basis. Since the two purposes of this survey were to determine for what purpose the items checked out were used, and where the patrons obtained their reference, the studies of Bernal and Herner previously mentioned are not directly applicable. They were more interested in evaluating literature research techniques and determining the relative usefulness of research information and reference services. However a comparison is made between the findings of this survey and those of Urquhart, Bernal, and Herner whenever it seemed applicable.

A total of 371 questionnaires were completed and of this number 173 were completed by undergraduates and 198 by graduates and faculty members. Thus, over 50% of the circulation of the library is to this latter group; 162 of the questionnaires were completed by individuals who had previously filled out a questionnaire.

The Roman numerals in the following sections correspond to the numbers in the questionnaire of Table 1. See Fig. 1 for a graphic presentation of responses.

Reason for Selecting Item

I. *For classroom work.* It is not surprising that undergraduates indicated that 47 percent of the materials they withdrew were for classroom work as compared with only 28 percent for the graduate-faculty group. To those who feel undergraduates do not engage in research the

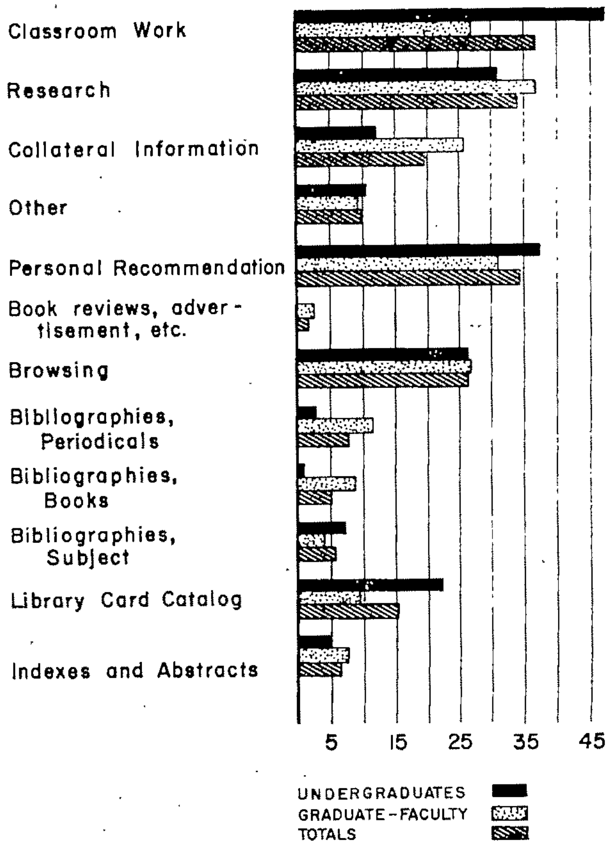


Fig. 1. Percentage for each response to the questionnaire for undergraduate and graduate faculty

total would be 77 percent. The responses between the three periods decreased 5 percent and then 10 percent as the semester progressed with the undergraduate group, whereas with the faculty-graduate group this reason decreased by 8 percent and 10 percent. This may be due to the fact of the individual definitions of research: to the undergraduate the completion of a semester paper implied research, whereas to the graduate student it meant class work.

II. *Applicable to research project or thesis.* 30 percent of the undergraduates and 37 percent of the graduate-faculty group gave as their reason for selecting an item that it assisted them in their research. The significance here may be that one-third of the circulation of the library is directly concerned with the research projects of its patrons. In both groups this reason for selecting an item increased during the three periods, 5 percent and 10 percent increases for the undergraduate, 10 percent and 16 percent increases for the graduate-faculty group.

III. *Information allied to the major field of interest.* In the first period two of every five items checked out by the graduate-faculty group were for collateral reading in their major field. By the next period the ratio had dropped to one to five and in the last period to less than one to six. The literature borrowed from the Science Museum Library was used mainly for "Technical Development Work," and "General Information" purposes. If these terms can be interpreted to mean that the materials drawn out were for the same purposes as referred to in this questionnaire, the average figure of 20 percent found in this survey is not inconsistent with that of Urquhart's. Whether this gradual decrease in the amount of materials drawn out of the library for collateral reading is an indication of the research study habits of the patrons cannot of course be definitively concluded, but it does reveal a periodicity which parallels the work ordinarily experienced during a semester.

The undergraduate group did not show much variance between the three periods. In the first period this reason constituted 14 percent of the responses and 10 percent in the last period.

IV. *For reasons other than I, II, or III.* This response was elicited approximately 10 percent of the time during the three periods by both groups. Although no statistical record was kept of the subject matter circulated, it was clear in this case that the ownership of an automobile, hi-fi set, or television set was the reason behind search for materials in this category in many instances.

Summary. Ten percent of the circulation was for reasons not connected with either research or class work. The graduate-faculty group charged out items in increasing amounts during the semester for class work and because the items were applicable to their research program. For this same group, collateral information to research or the major field of interest declined dramatically as a reason for selecting an item as the semester progressed. Sixty-two percent of the circulation of the graduate-faculty group were for reasons II and III. While increasing numbers of the graduate-faculty groups obtained materials for classroom work, the undergraduates gave this reason fewer times toward the end of the semester.

Source of Reference

V, VI, VII. *Personal recommendation.* This was clearly the most important reference source for both the undergraduates and graduate-faculty group. Interestingly, no graduate or faculty member and only four of nineteen undergraduates gave this source when they indicated reason 4 for selecting an item. As the semester advanced, this verbal source became increasingly important except for items selected for research by the graduate-faculty group. The undergraduates, on the other hand, indicated that at the beginning of the semester—two out of five instances—discussions with members of their department were the means of becoming acquainted with the item; by the last period the ratio had risen to two out of three. Two reasons might be given: the contact between individuals becomes more friendly over a period of time so that more discussions take place, and, as the semester advances, the student becomes more familiar with the subject matter he is studying to permit him to ask questions about areas he does not understand or he wants to investigate further. Urquhart's study shows that verbal recommendation ranked third as a source of reference constituting 16 percent of all circulated materials. Similarly, Bernal found this means to be the third in rank as a source of reference although the percentage was slightly lower (14 percent). Herner's figures are not directly comparable. Herner found that 50 perecnt of the information of the group he interviewed were obtained through verbal sources. But of the sources of information from literature that are "indirect sources of information," the estimate of 19 percent was given for personal recommendations as a means for learning about materials. However, the scientists at Johns Hopkins listed personal recommendations as the most important "indirect source" of information with the School of Engineering faculty ranking personal recommendation as second.

Verbal recommendations from other than members of a department or project were indicated relatively few times, 7 percent for professional contacts and 3 percent for contacts other than those connected with the patron's work. Since there were so few responses for these two "sources" and because they were so scattered, no particular trend or significance statistically was observable. However, in six of the ten instances, when the patron indicated the personal source of information was someone other than a member of his department or project or one of his own profession, the source was the librarian.

VIII. *Through advertisement, book announcement, or review.* This was clearly the least important of the sources of reference. No undergraduate gave this as a source. Of the graduate-faculty group, this source of reference was given five times.

The Engineering Library distributed only one accession list during this period. Herner found that the most appreciated library reference service was the publishing of accession and selected reading lists, but he does not give any statistics as to their relative use. The Science Museum Library survey showed accession lists to be the least important as a source of information. The Agriculture librarians at the University of Wisconsin report a considerable response from the distribution of their "new book list." They attribute this success to regular distribution of the list and the system of circulating the new books. Whether this source of reference is truly not considered a useful source by the patrons of the Engineering Library, or whether other factors are involved, needs further study.

There is evidence that faculty members do read book reviews and publishers' announcements because of their requests for purchase. Book announcements and reviews were seemingly of little use to the John Hopkins' scientists as they listed this source last. Neither of the British studies gave this as a possible source.

IX. *Browsing through the library collection.* One of every four items was learned about through browsing. It ranked second to personal recommendation. For the graduate-faculty group this ratio was relatively consistent throughout the period under observation. The undergraduates, however, spent less and less time browsing toward the end of the semester. Approximately one out of four items found through browsing was used as a source of information for the major field of interest, although the graduate-faculty group indicated that 19 percent of the materials they found in this manner were applicable to their research.

Since all the items charged out at the Science Museum Library in Urquhart's study were solicited by mail, browsing could not be considered as a source of reference. In only one of the questionnaires prepared by Scates and Yeomans (*4*) was this possibility for learning about graphic materials brought out. The question was so framed that it cannot be compared with the situation under study here.

X, XI, XII. *Through bibliographies.* Only one undergraduate indicated that his source of reference was a bibliography in a periodical. From counts made of individuals examining current periodicals, few undergraduates spend time reading the periodicals available in the Engineering Library; consequently, it is not surprising that undergraduates did not give this as an important means for learning about literature. Four students gave their source of information as a bibliography in a book. Because of the wording of response 9, there is no way of knowing what type of bibliography the 12 students used, but from the information obtained from several interviews, the bibliographies, in most instances, were those prepared by their instructors. These responses were too few and too scattered to show any trends.

The graduate-faculty group indicated that their source of information was a bibliography in one out of four instances. Of the total 24 percent, 10 percent were from books, 12 percent from periodicals, and 2 percent from especially prepared bibliographies. Bibliographies found in periodicals became an increasingly important source of reference toward the end of the semester, especially for items applicable to the research or thesis of the patron.

References cited in literature were the most frequent source of information for the individuals questioned in the British studies (37 percent by Bernal and 38 percent by Urquhart). The Johns Hopkins' scientists gave one-third their votes to this means (14 percent for bibliographies and 19 percent for books and papers).

XIII. *Subject heading in the card catalog.* The undergraduates found one-fifth of their materials through the subject card catalog, and it was used almost entirely for finding material for classwork and for their research. It was consulted only a few times for finding general information and only once when the patron was looking for material not connected with his school work. The use of the subject card catalog increased as the semester progressed.

The graduate-faculty group specified the subject card catalog as a source in only 9 percent of the cases. This total however is misleading: during the first period 16 percent of the class work items, 8 percent of the research items, and 22 percent of the items for general interest were found through the card catalog. By the third period, however, only one person in this group admitted using the card catalog—and he used it to get allied information for his thesis. Apparently the more advanced the work of a graduate student or faculty member, the less he depends upon the subject card catalog.

Herner found that the Johns Hopkins' scientists used the card catalog in about the same percentage as the graduate-faculty group did in this study. Subject card catalogs were not listed as a source of reference in the British studies.

XIV. *Indexes and abstracts.* From the responses on the questionnaires, indexes and abstracts were used very little as a means of becoming acquainted with literature. However, an actual count of individuals using the abstracts and indexes indicates that they are used as a reference source more frequently than this survey shows (*5*). Indexes and abstracts deal almost entirely with periodicals. Periodical references, when found through this means, are read in the library and only the especially useful or pertinent item is charged out. In all three of the studies previously mentioned, indexes and abstracts ranked near the top as the most useful reference source.

Summary. Decidedly, verbal recommendation plays an important role in informing the patrons that there is literature of interest to them in the Engineering Library. Over one-third of the items charged out were recommended to the patron by someone. Over one-fourth of the items circulated were learned about through browsing in the library collection. Bibliographic references are a relatively unimportant device for the undergraduate, whereas one-fourth of the items charged out by the graduate-faculty group were discovered through this means. Book reviews and accession lists do not seem to be influential in increasing the library's circulation. One-fifth of the items charged out by the undergraduates were discovered through the subject card catalog. The subject card catalog seems to play a relatively minor role for the graduates and faculty members as their work advances.

• Conclusions

The data obtained through this survey cannot be used to make generalizations for all libraries or even for college engineering libraries. Two important failings are observable in this method of survey: (*a*) that only certain aspects of the total library operation were under observation; and (*b*) a statistical method, while an important means for interpretation of situations, can never convey individual differences that are so important where human preferences are involved. At best a statistical approach to problems can reveal "general trends" and indications of directions for action. In this respect this survey has emphasized some well-known, but often not considered facts about library practice.

1. There were as many graduates and faculty members using the library's non-reserve collection as undergraduates although statistically they are a very much smaller group in the College of Engineering.

2. The most important source of reference to materials in the library was through *verbal recommendation*. This has important implications for the librarian if he is to take part in "information gathering habits" of his patrons. First, the librarian must be acquainted with the subject fields of his patrons. This is perhaps obvious, but it is not unimportant because it is obvious. A librarian cannot discuss the literature of a subject field without knowing something about the subject field. Second, the librarian must be acquainted with his patrons as individuals. One ordinarily does not discuss even abstract and impersonal matters with a stranger.

3. Because of the relative importance of browsing in the library, the physical arrangement of the materials undoubtedly is an important factor. In other words, the "availability" of the library collection determines how it will be used.

4. Perhaps more important than any "generalization" that has comes from this survey has been the personal insight that was given to the librarians at the Engineering Library with regard to their own work in relation to the work of the staff and students of the College of Engineering. It has brought to attention and emphasized the interdependence of teaching and research with functioning of the library.

References

1. Herner, S., Information gathering habits of workers in pure and applied science, *Industrial and Engineering Chemistry*, 46: 228–236 (1954).
2. Bernal, J. D., *Preliminary analysis of pilot questionnaire on the use of scientific literature*, in *The Royal Society Scientific Information Conference, 21 June–2 July, 1948, Report*, The Royal Society, London, 1948, pp. 589–637.
3. Urquhart, D. J., The distribution and use of scientific and technical information, in *The Royal Society Scientific Information Conference, 21 June–2 July, 1948, Report, The Royal Society*, London, 1948, pp. 408–419. Reprinted from *Journal of Documentation*, 3: 222–231.
4. Scates, D. E., and A. V. Yeomans, *Activities of Employed Scientists and Engineers for Keeping Currently Informed in Their Fields of Work*, American Council on Education, Washington, D. C., 1950.
5. Schubrin, A. W., and C. A. Schauer, unpublished report, May 31, 1956.

Letters to the Editor

Dear Sir:

More concerning full names vs. initials as per letters to the editor [*American Documentation,* January 1967] re: W. T. Brandhorst. My approach is, why worry? Establish a standard similar to the military, i.e., "sure everybody has a name, buddy, but here you are just a number." So, what better identifier for Joe Author than his Social Security number? Editors will not accept for publication any document where the author has not included his SS#. (Hmm! better make that *Social Security* number). Those who want to file from now on, file by number, x-ref to a desk-top author-number list, and who cares how you sign your name!

Daniel M. Simms
Mobil Oil Corp.
Dallas, Texas

Dear Sir:

In the paper presented by Barbara A. Montague to the 1964 meeting of the ADI, it is claimed that three systems of indexing were compared, *A* and *B* being "co-ordinate indexes," and *C* a "classification index." This exercise seems to me to be one of the most carefully constructed and executed of any that I have seen in this field, and its results, I find, are not only completely convincing but also, unlike many other such tests, actually relevant to the task of information retrieval in real life.

It is the more the pity, therefore, that Miss Montague was not more explicit in describing her three systems, because her conclusions are liable to give an utterly false impression. What made system *A* superior to system *B* were (1) its system of vocabulary control, (2) its ability to provide for generic search, and (3) its superior use of roles. As to (2), I should like to emphasize Miss Montague's own point, that "the main factor responsible for irrelevance in system *A* occurred in one question *for which selective generic class was not available,* and the classes which had to be searched *included concepts unrelated to the question"* [my italics].

Now system *C* is described only as a "subject index with one or two cross references on abstract cards." This corresponded to no "classification index" that I have ever heard of. A "classification index." I should have thought, was the *alphabetical* list of terms, with their class numbers, that one finds at the end of the classification schedules, at least in all the schemes I know. Yet, in the conclusions, this apparently alphabetical list, so primitive that it has only one or two cross references, has become "a classification system." The next thing we shall have is the claim becoming accepted that Miss Montague has "proved" that co-ordinate indexing is superior to classification for information retrieval. What, I should like to ask, is system *A*'s ability to provide for generic search if it is not classification? What does Miss Montague imagine a faceted classification to be if it is not a controlled vocabulary, with generic structure, with precise roles (these are the facets), and also with selective generic classes for *all* its terms? Every co-ordinate index that departs from mere alphabetical listing of terms, and introduces generic structures and roles, has become a classification system.

I should not write with such emphasis if I did not estimate highly the value of Miss Montague's work. I agree with her views on deep indexing and the use of roles; I agree with her views on links, except that I wonder whether a very selective use of links, by a skilled indexer—and "when in doubt, leave it out"—might not sometimes improve performance; and I would emphatically endorse the implicit conclusion that the system which costs more at input but less at output is the better. I should grieve to see a paper of this quality cited as evidence for claims which it actually proves to be false.

D. J. Foskett
University of Ibadan
Ibadan, Nigeria

Dear Sir:

I appreciated the opportunity to read Professor Foskett's penetrating analysis of my paper, "Testing, Comparison, and Evaluation of Recall, Relevance, and Cost of Coordinate Indexing with Links and Roles," which appeared in the July 1965 issue of *American Documentation.*

Professor Foskett fears that readers of this paper may misinterpret my statement that the coordinate index, System A, performed better than System C, which was a classification system, and extrapolate this to mean that coordinate indexing is superior to all classification systems. Indeed, this was not my intention and, being aware of such potential misinterpretation, I carefully worded my statement as follows in the **Summary and Conclusions:**

> The comparison of *a* coordinate index with *a* classification system shows that, *for the two systems tested,* coordinate indexing provides faster searching and retrieves more relevant references, and that the cost of coordinate indexing is higher at input and less for searching.

The objective of the test reported in the paper was to evaluate the relative effectiveness of the two systems. System C, primitive though it may have been, was the practical search tool actively used by the patent attorneys themselves at that time, and the success of its performance depended in large part on the familiarity in depth of the body of art by the users. It was concluded from this test that the performance of the coordinate index System A justified the cost of input and was superior to the classification system in use at that time.

It is obvious from Professor Foskett's letter that we share common views in improving the understanding of a wide variety of approaches in use today for the storage and retrieval of information. It is this writer's hope that opportunity will arise in the future for personal discussion with Professor Foskett in these areas of mutual concern.

Barbara A. Montague
Information Systems Division
E. I. du Pont de Nemours & Co.
Wilmington, Delaware

Dear Sir:

"Computer" and "Boolean" were magic words for many in documentation during the fifties. Since then we have learned a good deal more about what electronics and symbolic logic can and cannot do for documentation. Now "behavioral" and "psychological" seem to have become magic words for some who write about documentation problems. A striking example appeared in the *American Psychologist*, November 1966:

> Many of the missions in our society have technological goals, such as sending a manned space vehicle to the moon or producing efficient thermonuclear power. The information science mission, however, is to facilitate a peculiarly human intellectual behavior, namely, research itself. Because of this essential difference, the criteria for evaluation of an information system cannot be specified as technological ones. They must be specified as behavioral ones. The development of criterion measures to instruct the designers and operators of information systems (whether journal editors, planners of computer-based reference-retrieval systems, or whomever) about what they should try to accomplish and how to measure their success thus becomes a task for psychological research (*1*).

The last sentence of this passage is noteworthy. If it had ended "becomes a task which can be helped by psychological research," it would be plausible. But it would also be plausible if it had ended "becomes a task which can be helped by history of science research" or "philosophy of science research," or "documentation research," or "library science research," etc.

The sentence as it stands claims a central, and apparently self-sufficient, role for psychology in developing criteria for evaluation of information systems. The paper from which the sentence is taken provides no support for that claim. Anyone who finds the quoted passage plausible should note in it the ambiguous occurrence of "behavioral." Specifically, psychology as a "behavioral science" may take all human (and animal) behavior as its subject. But it does not follow that psychological inquiry can answer all questions about that subject—all "behavioral" questions. For instance, driving an automobile is behavior. But psychological research cannot instruct highway engineers "about what they should try to accomplish and how to measure their success"—though it can help in developing such instructions.

The point made in the closing four sentences of the preceding paragraph is generally applicable to attempts to apply the behavioral sciences in documentation.

Reference

1. PARKER, E. B., and W. J. PAISLEY, Research for Psychologists at the Interface of the Scientist and his Information System, *American Psychologist*, 21: 1069 (1966).

JOHN O'CONNOR
Center for the Information Sciences
Lehigh University

Ed. note: This note is a by-product of the author's work on information retrieval, which is sponsored by the Information Systems Branch, Office of Naval Research (Contract Nonr N00014-66-C0096). Reproduction in whole or in part permitted for any purpose of the United States Government.

Book Reviews

3/67–1R **Data-Retrieval by Computer: A Critical Survey.** 1966. Asa Kasher. The Hebrew University, Jerusalem, Israel. 76 pp.

"critical": 1. Inclined to criticize, esp. unfavorably; captious; censorious
2. Exercising, or involving, careful judgment; exact; nicely judicious

This report defines "data retrieval" in a very unfamiliar but completely valid way, so as to encompass both data base systems and text processing systems. But it devotes most of its critical analysis to what have generally been called "question answering systems," which makes the use of the term "data retrieval" particularly strange. That is, it is concerned with a critical evaluation of a number of computer programs which accept natural language sentences as input and from them generate logically acceptable output. Such systems involve a combination of problems in language data processing, file searching, and logical calculus. Table 1 presents a chronological listing of those which were considered in this report. For each, there is a brief indication of the approach which the system seems to have adopted in each of the three areas of analytical processing.

The prospective reader will find it advisable first to read "Answering English Questions by Computer, A Survey," R. F. Simmons, *Comm. ACM* 8 (1): 53–70, Jan. 1965, since much of this report not only refers to Simmons' survey, but apparently has the intent of negating his evaluations as well. Specifically, the Kasher report raises three primary critical problems and a number of subsidiary ones. It claims that:

1. None of the systems examined has an adequate means for resolving linguistic ambiguity—of syntax, meaning, or context—even though the descriptive reports about them typically imply differently.

2. None of the systems has an adequate decision method for handling logical consistency in the input data, since they all ignore some basic theoretical problems in logic.

3. None of the systems has an adequate definition of what constitute "questions" and "answers" since at best they represent explications of specific types of questions (or answers) on specific subjects.

Traditionally, a critical review is "critical" in sense 2 above. This one is more than critical, even in sense 1; it is pejorative, sprinkled liberally with phrases such as "methodological errors . . . widespread misconception . . . serious flaws which indicate its fallaciousness . . . claims which attempt to disregard problems . . . devoid of practicality in an essential way . . . ignoring significant . . . theoretical results . . . results are almost trivial . . . absurdity of his claim."

The view of the report as a whole is summed up on pages 66 and 67 as follows: "There are those who consider that . . . question-answering deals only trivially with a trivial sub-set of English. . . . The faults are far more serious, in that they stem from grave difficulties of principle." On the basis of its critical analysis, the report concludes that ". . . the only hope for success in the near future is in well-structured data-based systems, having a special internal structure appropriate to a specific field, a reliable technical language, and a competent inference-mechanism, the latter taking account of the special internal structure and based, as far as possible, on non-classical calculi."

The report's conclusion seems eminently reasonable, but one wonders in what way the question-answering systems which are attacked in fact depart from it. As the report's analysis demonstrates, there is no essential difference between the "data-based" systems and text-based ones (under the proviso, which the report so strangely adopts, of English language input to each); each of the systems indeed embodies a special internal structure, presumably appropriate to its specific field; each utilizes a specific technical language which, while it tries to approximate English, presumably is reliable; and most of them include an inference mechanism, usually a variant of the predicate calculus, each presumably based on its special internal structure (although its "competence" may be subject to question). Why then the need to attack what is essentially a straw man—namely, that the designers of these systems have the hope and desire of making a positive step toward a formalized means of handling natural language? Admittedly, the descriptions of desires have become an almost ubiquitous part of what should simply be reports of results. But the informed reader of these reports has by this time surely learned to filter them out and evaluate the actual results for what they are—operating models, more or less illuminating as realizations of prospective theories, which also provide the means for applying those theories to a relatively large number of examples. The pity is that someone so capable of "critical" review did not also attempt to define and evaluate the *positive* contributions which each of the projects he analyzed has made to the goal which the report itself defines as the "hope of success."

As a step toward such an evaluation, the techniques each system uses can be roughly classified into the three categories defined previously—language data processing, file searching, and data reduction (including logical processing). Because it is in fact the *combination* of problems (and techniques for solving them) which the systems have tried to handle, it seems to be worthwhile to evaluate the effective contributions of each. To this end, it would be useful to analyze each of the systems in terms of the following successive areas of complexity, and perhaps quantify them by the size of the *tables* involved in each project:

1. The length of strings which can be accepted, as a group for analysis.
2. The vocabulary, quantified by the number of prestored terms the program handles.
3. The semantic ambiguity, quantified by the average number of separate meanings a representative term may have.
4. The syntactic ambiguity, quantified by the average number of syntactic roles for terms.
5. The richness of syntactic patterns, quantified by the number of defined sentence patterns.
6. The magnitude of the retrieval task, quantified by the number of stored sentences.
7. The complexity of the measurement of degree of match, quantified by the number of alternations of logical operators.
8. The complexity of the logical analysis, quantified by the number of stored sentences which can be simultaneously considered.

The characterization of the complex tasks of language analysis and logical inference in such simple terms as "size of stored tables" is admittedly a gross over-simplification, which certainly doesn't even begin to recognize the theoretical problems involved. However, as Kasher in his report rightly points out, at best the systems analyzed are concerned with the development of "technique" and do not

TABLE 1

Person and system	Journal* and date	Language data processing	Searching	Data reduction
Phillips "ORACLE"	1960	One syntactic pattern (SVO, Time, Place)	Simple Matching on words in syntactic pattern	
Sable "IDL"	C-ACM, Jan. 1962	Generic Relations		
Salton	C-ACM, Feb. 1962	Tree-Structures		
Householder "Auto Lang Anal"	1962	Roget's Codes	Correlation of terms	
Simmons "Synthex"	AD, Jan. 1963		Complete Index	
Harrah	Communication: A logical model, 1963			Predicate Calculus (definition of questions and answers)
Green, Clonsky "BASEBALL"	1963		List structures	Simple counting
Simmons "Protosynthex"	1963			
Lindsay "SADSAM"	1963	Basic English words (1413) and syntactic classes. Predictive analysis. Specific syntactic class defined formats	List structures	Tree structure for defined relations
Doyle	Nov. 1963	Statistical properties		
Cooper Fact Retrieval	J-ACM, Apr. 1964	Sublanguage of English. Trans. by *grammatical classes* into "logical" equivalents. Specific algorithm for constituent structure languages which steps thru grammatical classes	All possible subsets of up to three stored data sentences, each of which contains at least one basic term from the question	Aristotelian logic
Salton "SMART"	June 1964	Specific syntactic structures including "semantic" classes)	Matching terms with weights in structure	Statistical Tree-structure
Simmons "Synthex"	AD, Jan. 1963		Complete Index	
Bobrow "STUDENT"	AD 604730, Sept. 1964	Kernel sentences (20 different formats) Conversion routines	Simple table look-up	Arithmetic for simultaneous linear equations
Raphael "SIR"	1964	20 fixed formats		Set relationships Simple arithmetic
Cooper	1964			Syllogisms and Propositional Calculi
Darlington	1964			Predicate Calculus
Thompson "DEACON"	1964	Word classes (functions, lists, attributes, modifiers)	List structures	
Kirsch, et al. "PLM"	1964	Descriptions of pictures Phrase-structure		1st order predicate calculus
Black "SQA"	1964	Fixed Formats	Exact match on words and structure	Rules of inference

* Abbreviations: C-ACM: Communications of the Association for Computing Machinery

really provide any real insight into the theory of language or logic. The issue then is the power or effectiveness and this can be measured by the size of task which can be encompassed, particularly if this can be related to the processing times and equipment costs.

ROBERT M. HAYES
Institute of Library Research
University of California

3/67–2R **National Library of Medicine Current Catalog.** Volume 1. Jan. 1–14, 1966. U.S. Public Health Service, Washington, D. C. Biweekly, cumulated quarterly from the first of the current year; annual cumulation casebound. Supersedes *National Library of Medicine Catalog.* Sold by the Superintendent of Documents: 1966 price, $15.00 a year ($20.00 foreign) with annual cumulation also sold separately for $4.50; price increase to be announced.

Current Catalog is not likely to hold still long enough for description in the timeless or at least timely terms reviewers would like to apply to books. Having launched it as a somewhat tentative venture in using new techniques to meet old goals (its machine system is called an "interim module"), its mentors not only have had their own plans from the start for further development, but also have been actively seeking users' criticisms of both its purposes and its performance with the implication that these responses will affect its future.

Preparation for *Current Catalog* began during the period when Dr. Frank B. Rogers led the National Library of Medicine beyond a past of proud distinction to revolutionary advances in organizational status, physical facilities, and bibliographic services, which culminated in the remarkable production of the *Index Medicus* as one output of a computerized Medical Literature Analysis and Retrieval System (MEDLARS). Many people have guided the development of *Current Catalog,* some now gone from NLM (Samuel Lazerow, then Chief, Technical Services Division, and Irvin Weiss, a systems analyst), some still there (Scott Adams, Deputy Director), and others recently arrived (Dr. Martin M. Cummings, Director since 1964). Most directly responsible now are James P. Riley, Chief, Technical Services Division, and Emilie V. Wiggins, Head, Catalog Section.

A printed catalog of the books in the National Library of Medicine is not new. One of the earliest was a 454-page volume issued in 1872, followed by a three-volume edition in 1873–74. (The "first catalogue" was an 1840 manuscript "containing 23 unnumbered leaves and listing 130 titles," not printed, however, until a facsimile edition was published in 1961). The monumental *Index-Catalogue of the Library of the Surgeon General's Office,* whose sixty-one volumes issued in five series from 1880 to 1961 record on their title pages the evolution of official names from "Library of the Surgeon General's Office, U. S. Army," through "Army Medical Library" and "Armed Forces Medical Library," to the present Congressional designation as "National Library of Medicine" (under the Public Health Service), contained author and subject entries for periodical articles. Before the decision was implemented to finish listing those books with imprints through 1950 by issuing the three volumes of Series 5 (1959–61), the Library began publishing as a new work, supplementary to the *Library of Congress Catalog,* an annual volume recording its book cataloging of the previous year: an April–December 1948 volume and a 1949 volume were issued before the regular annual series of the *National Library of Medicine Catalog* (earlier titles: *Army Medical Library Catalog* and *Armed Forces Medical Library Catalog*) began in 1950, with superseding five-year cumulations following for 1950–54 and 1955–59, and a six-year cumulation for 1960–65 soon to appear as an end to that series. In addition, from 1960 through 1965 a monthly list of catalog main entries for selected U. S. books and periodical titles appeared without cumulation at the end of each *Index Medicus* issue, entitled "Recent United States Publications."

The *Current Catalog* continues the record of NLM book cataloging from January 1966. As an NLM published catalog, it is new in that it attempts rapid reporting of cataloging data through biweekly issues, cumulated each quarter ("Cumulative Listing") for the entire current year to date, with a bound cumulation to appear annually and perhaps larger cumulations to follow later (although no announcement has been made on the possibility of larger cumulations). The most similar existing service is *The National Union Catalog,* which the Library of Congress issues monthly with quarterly cumulations, but which does not provide the same degree of currency in reporting or cumulating as *Current Catalog* promises, nor the same complete listing of added entries. One may assume, further, that neither *NUC* nor any other published book catalog covers biomedical books as comprehensively as does *Current Catalog.*

Current Catalog is new also in that its data are first machine encoded, next manipulated by a computer to alphabetize them under all desired entries with varying parts of the data appearing under the several types of entry, then composed by the computer in column and page format, and, finally, by GRACE (Graphic Arts Composing Equipment), NLM's computer-driven, high speed Photon 900, reproduced automatically in high quality type fonts on photographic negatives ready for offset printing. The process, similar to that by which *Index Medicus* and other recurring bibliographies are prepared, thus combines speed of editorial assembly and page composition with the readable appearance of traditional book type (upper and lower case, boldface and roman, serifs, and diacritical marks), although the space-saving six-point type is still a frustration to tired eyes. In addition, the machinable records have the potential for such varied new uses as computer compilation of demand bibliographies and distribution via punched cards, punched paper tape, or magnetic tape to other libraries for their use in producing catalog cards or in computer searching.

At the moment, however, NLM does not have programs ready for selective searching of the book cataloging data as it does for MEDLARS periodical indexing data (a cataloging record can be accessed only by citation number to make corrections), and it has not merged any of these book entries with its MEDLARS data file (pre-1964 plans called for beginning with a combined file, but the problems were too great).

The catalog entries in *Current Catalog* are the result of NLM's traditional book cataloging practices, which presently are not accompanied by subject cataloging in depth (assigning more than the average two or three subject headings in order to identify specific and multiple subject aspects useful in computer searching). Although the printed *Current Catalog* would not itself reflect such additional subject cataloging, the project surely must attempt someday to benefit from the opportunities which machine searching presents for more effective storage and retrieval of information from books than traditional subject cataloging allows. The theory and procedure of subject cataloging books in depth are full of problems, however, which require more than machines or increased manpower to solve.

The uses and stated purposes of *Current Catalog* are several: it is a comprehensive announcement list from which other libraries may select new acquisitions; it is a source of cataloging data to assist cataloging efforts in other libraries; it is a permanent reference tool for manual searching of the literature under names, titles, and subjects.

Current Catalog's attempts to publish cataloging data in time for use in acquisitions and cataloging activities of other libraries so far have been only partially successful. NLM has arranged to receive advance copies of domestic publications and by a tour of Europe by one of its staff members has made similar arrangements with European publishers, but these arrangements have not yet (January 1967) resulted in the rapid cataloging desired. Several libraries have reported recent comparisons of current imprints listed in their published accessions lists with those in all issues of *Current Catalog* through the corresponding date; the comparisons showed a large proportion of relevant new titles not listed in *Current Catalog.* While it is an excellent complementary selection tool, especially for government and foreign books and for Public Health Service and other government contractors' biomedical reports, it cannot yet be used as a medical library's primary selection source.

Delays have resulted at other points in the *Current Catalog* production process in addition to that of cataloging input. Most obvious is the delay in distribution after GRACE has readied the page negatives for the nongovernment printer, who usually gets the printed issues to the Government Printing Office mailers within four days. While each 1966 biweekly issue was intended to reach subscribers on the final "coverage" date then printed on the issue, this reviewer's library in common with others regularly was receiving its issues over two weeks later. The U. S. mail might be a bottleneck, but the GPO mail room seems the culprit of choice. NLM and some of the subscribers have recently toyed with the idea of sending advance photocopies as an experiment in reducing the delivery time, but no mass solution is yet indicated.

In addition, a proposed distribution of machinable cards or tapes prior to the printing of each issue may speed delivery to those libraries able to use such records. (A breakdown of GRACE in December and January delayed photographic copy of the first 1967 issue for eight days. The Library announced that it would arrange for back-up equipment to prevent similar future delays.) Speed in production is perhaps most notable in issuing the cumulations: the first annual cumulation is expected at this writing to be in the hands of the mailers by the first week in February, considerably sooner than the former annual catalog used to be.

An early change was made in page content to assist visual scanning for selection purposes, as well as to save space. At first, the full citation (e.g., author, title, edition, place, publisher, date, collation, series, other notes, tracings, call number, an in-house citation number, and price, when known) was reprinted from author through date plus the call number and citation number under all added entries (joint authors, editors, title, series, etc.), but beginning with the third quarter of 1966 issue of July 2–14 the added headings were entered as "see" references to the main entry heading without reprinting of any other part of the citation except the citation number (and in 1967 the citation number also will be omitted from these cross-references, so that a single cross-reference might suffice to cover more than one title or version). The resulting short block of print, it has been suggested, will signal added entries which can be skipped when scanning for selection purposes, although some few main entries are themselves equally short and must be watched for. The main entry heading, whether in the main entry itself or in a cross-reference, is in boldface as an added aid to scanning, but six-point boldface and roman type can very quickly blur into the indistinguishable.

The biweekly issues, 20 by 25.7 cm. in size and in two-column format, include main and added entries plus subject entries for persons or corporate bodies for imprints of the current or two preceding years, but not topical subject entries (although full tracings are shown under each main entry). Each biweekly issue has at the back a separate list of "Added Volumes" newly received and an alphabetical "Directory of Publishers" of books with their addresses and a sublisting in citation-number order of addresses of publishers of serials, all very useful for acquisitions purposes. Newly cataloged earlier imprints (except pre-1801 imprints and Americana) and, in a separate "Subject Section," full citations except for tracings and price under all topical subject headings (which NLM with some exceptions does not assign to books over twenty-five years old) are added to the quarterly cumulations (the pages are larger—23.5 by 29.5 cm.—and hold three columns). The quarterlies further differ from biweeklies in including cross-references from variant forms of names. The January–September 1966 quarterly, third for the year, contained a Subject Section of 374 pages and a Name (and title) Section of 457 pages.

Recent NLM discussions with subscribers have involved suggestions that topical subject entries would be useful for selection and reference purposes even in the biweekly issues, and, according to the *National Library of Medicine News* 21 (12): 4–5, Dec. 1966, a subject listing will be included in the biweekly issues, to be enlarged to the same three-column format as the quarterlies, beginning January 2, 1967 (the date of the issue will then be the closing date for corrections and changes), with an accompanying price increase to be announced. Some subscribers have suggested adding tracings for name cross-references and including these in the biweekly issues for catalogers' use. Some have proposed adding union catalog holdings information, a separate supplement of new serial titles, and Library of Congress subject headings and classification numbers.

The new Anglo-American cataloging code is apparently being followed more by NLM than it will be by LC (since July 1966, NLM has been using a draft copy of the new code, although it plans to wait for future machine assistance before making certain large changes, such as dropping "U. S." at the beginning of every entry for the National Library of Medicine). It would be highly useful if the several national distributors of centralized or shared cataloging could more consistently accept and apply rules of entry and description. It is one thing to ensure finding of a book by putting multiple entries and cross-references in a card, printed, or computerized catalog, but it is not satisfactory to have a variety of possible main entries when only one entry is to be listed in selective bibliographies, in a library's acquisitions files, or on temporary catalog slips. (A curious example of variation in main entry is a comparison of *New Serial Titles'* corporate main entry, "U.S. National Library of Medicine. Current catalog" with NLM's title main entry, "National Library of Medicine current catalog." The former follows past ALA rules, while the latter demonstrates NLM's preference for a title main entry when a corporate name comes first in the title but would have to be altered to follow ALA rules (in this case by adding "U. S.," although the Anglo-American code will drop this particular "U. S."). Otherwise, NLM too has followed ALA rules for corporate entry of periodicals. (It so happens that the title-page form of title entry for periodicals is usually the more useful to readers familiar only with the citations used in indexing and abstracting services and in the periodical literature itself, and one might conclude that the new Anglo-American code, which retains the corporate entry rule for periodical titles like *Journal of the American Medical Association,* does not wholly suffice as a basis of agreement, although it will be the best code libraries have ever had.) Complete consistency in main entry is impossible, needless to say, but a concerted attempt by major libraries to accept a specific main entry for a specific book or serial might be found useful upon reevaluation of the problems which are supposed to prevent such unanimity. The results would assist readers in discovering whether a particular book is held by a library, as well as reduce the expensive adaptations of shared cataloging which go on in library after library.

The National Library of Medicine has been seeking suggestions not only from subscribers to *Current Catalog,* but also from independent systems analysts (the Auerbach Corporation), who are helping to design improvements in many aspects of NLM's data storage and retrieval activities, including MEDLARS, the *Current Catalog,* acquisitions and serials records, and ways of handling graphic images. *Current Catalog* may be expected to change even more as new methods are found to meet old and new goals at a price the smaller libraries can afford to pay (and the original price was an outright bargain).

Current Catalog is an essential part of the bibliographic apparatus, which every library seriously interested in medicine or the biomedical sciences must have and use.

Stanley D. Truelson, Jr.
Yale Medical Library

3/67–3R **Scientific Management of Library Operations.** 1966. Richard M. Dougherty and Fred J. Heinritz. The Scarecrow Press, Inc., New York, 252 pp.

It is a pleasure to commend this volume to the readers of *American Documentation.* As a long-time admirer of the work of Ralph Shaw, I had come in recent years to think that the down-to-earth, practical aspects of work analysis in relation to library objectives were being neglected in favor of the more sophisticated "systems-analysis" approach associated with automation. It was a double pleasure, therefore, to find this handbook type of presentation, effectively combining general statements of principle with precise advice and directions on how to go about making management studies. The volume has the further advantage of a judicious list of references for further study, including such classics in the field as those by Frederick W. Taylor and Lyndall Urwick, as well as those as recent as 1964. The work of some 258 pages, including index and numerous charts and illustrations, has thirteen chapters. Key word or phrase subheadings, both in the text and in the table of contents, facilitate quick reference and orientation to what is to follow. It is possible, accordingly, for the prospective user of the volume to determine almost at a glance what might be relevant to a particular problem.

The chapter headings will in themselves give something of the flavor of the book: I. Scientific Management: What It Is and Is Not; II. Making a Management Study; III. The Flow Process Chart, Flow Diagram and Block Diagram; IV. Decision Flow Charting; V. Operations Analysis Including Some Principles of Motion Economy; VI. Forms: Their Analysis, Control and Design; VII. Time Study; VIII. Sampling; IX; Aids to Computation; X. Cost; XI. Performance Standards and Control; XII. Study of a Circulation System—The Present Method; XIII. Study of Circulation System—The Proposed Method.

Moving from these relatively general topical headings, the text itself is presented in a direct and lively style, but at the same time with the kind of restraint that is all the more persuasive because one feels that he is being talked with rather than being talked at. Examples of this approach can be found on almost every page. Examples: on page 16: "Libraries share with the multitude of other governmental and public service organizations in any community the responsibility for giving the taxpayer a maximum return of service for each dollar invested." A little later (page 17), "In addition to improving routine efficiency, management is a useful tool of library personnel management and financial administration. Work analysis is the key to modern job classification. Only when we have ascertained of what the job consists, and what level of productivity we may reasonably expect of the person performing it, are we able to define intelligently what sort of innate ability and special training are necessary for its performance."

The rhetorical question and its variations are used fairly frequently, but again, for this type of volume, effectively. On page 17 for example, "This argument against the efficacy of scientific management in libraries disregards the very substantial part of library work that consists of repetitive, mechanical routines that lend themselves readily to quantitative analysis. In terms of total hours required for performance, the largest bulk of library work—perhaps as much as 70 to 90% of all current library tasks—consists of such routines."

The practical approach is stressed throughout. Under the heading "Selecting an Area for Study," one finds the subheading, "Frequently Performed Jobs," and then in the text, "Since the time, money, or energy are not available to study everything and everybody, our efforts must be concentrated upon areas that are likely to yield the highest return for our study investment. The more frequently an operation is performed, the better a candidate it is for analysis. The reason for this is that even if we are able through improvements to save only a small amount of time each time the operation is performed, this saving multiplied by the high frequency makes the total time saved substantial." Although the statement would be obvious and self-evident to the experienced person, it is precisely this kind of information which should be part of the education of all librarians, whether or not they themselves are to be engaged in this type of management work.

A further convenience of the volume is the extent to which background information necessary in applying these techniques to library operations is included. In the chapter on sampling, for example, one finds: "The closer to 100% certainty that an investigator demands that his sample approach, the larger it will have to be. He must therefore begin by making two decisions as to how reliable an answer he needs or desires. The most common practice is to use a 95% confidence level. This means that the sampler can be confident that his random observations will represent the facts 95% of the time. It also means that 5% of the time they will not. . . . The most common confidence level is 99%. Since a 99% confidence level is seldom necessary in management practice, and since the 4% increase in certainty may require a substantial increase in sample size, it is recommended that it be used sparingly."

The examples given above illustrate another characteristic of the presentation throughout the volume, namely, the self-confidence with which the advice, conclusions, and recommendations are given. While omission of qualifying phrases may result in an oversimplification of the problems inherent in making management studies, the direct, concise, and lucid text is to be preferred for the purposes for which the volume was written.

While librarians and others who have been following the literature of this field through the years will find little that is new, Daugherty and Heinritz have made a real contribution to the literature of librarianship by bringing so much information from past and current practice together in one convenient volume.

Richard H. Logsdon
Director
Columbia University Libraries

ADI Chapters and Secretaries

CENTRAL OHIO CHAPTER
Mrs. Arleen Sommerville
Chemical Abstracts Service
Ohio State University
Columbus, Ohio 43210
614-293-8933

CHICAGO CHAPTER
Miss Patricia Llewellen
IIT Research Institute
10 West 35th Street
Chicago, Illinois 60616
312-225-9630

DELAWARE VALLEY CHAPTER
Miss Marilyn Leasure
Technical Library—Louviers
E. I. du Pont de Nemours & Co.
Wilmington, Delaware 19898
302-366-4242

INDIANA CHAPTER
Mr. Asa N. Stevens
6157 E. St. Joseph Street
Indianapolis, Indiana 46219
317-357-6460

LOS ANGELES CHAPTER
Sister Mary Lucille
Dean, School of Library Science
Immaculate Heart College
2021 N. Western Avenue
Los Angeles, California 90027
213-462-1301, ext. 297

METROPOLITAN NEW YORK CHAPTER
Miss Betty Jean Dougherty
Port of New York Authority
111 8th Avenue
New York, New York 10011
212-620-7000

NEW ENGLAND CHAPTER
Miss Virginia Valeri
Arthur D. Little, Inc.
15 Acorn Park
Cambridge, Massachusetts 02140
617-864-5770

NORTHERN OHIO CHAPTER
Miss Helen Skowronska
Sherwin-Williams Co.
P.O. Box 6027
Cleveland, Ohio 44101
216-TO1-7000

PITTSBURGH CHAPTER
Mr. James Brandt
ALCOA Research Laboratories
Box 772
New Kensington, Pennsylvania 15068
412-337-6541

POTOMAC VALLEY CHAPTER
Mrs. Joan Mavity
Herner & Co.
2431 K Street, N.W.
Washington, D.C. 20037
202-965-3100

SAN FRANCISCO CHAPTER
Mrs. Anne Raphael
176 Osage Avenue
Los Altos, California 94022
415-323-6138

SOUTHERN OHIO CHAPTER
Mrs. Esther Norton
6061 Crittenden Drive
Cincinnati, Ohio 45244
513-684-3111

SOUTH TEXAS CHAPTER
Mr. Doug Yauger
7107 Augustine
Houston, Texas 77036
713-PR4-1269

UPSTATE NEW YORK CHAPTER
Mrs. Pauline Atherton
School of Library Science
Syracuse University
Syracuse, New York 13210
315-476-5571, ext. 3823

When 300,000 scientific and technological articles are indexed each year--and each is indexed to a depth of approximately 50 entries -- that's subject indexing at its finest. And it's coming in the Fall of 1967 in the PERMUTERM™ SUBJECT INDEX to Science and Technology.

Librarians are asking for it. And *ISI* is delivering it—the *PERMUTERM SUBJECT INDEX 1966* to Science and Technology. *PSI*™ will quickly locate articles for you on the specific or generic subjects you are interested in through this daring new concept in subject indexing. Using the computer, like a computer was meant to be used, *ISI's* exclusive Permuterm programs index the average article to a depth of approximately 50 entries. Sample formats and details are available now. Books will be delivered in the Fall, 1967. Wait for the *PERMUTERM SUBJECT INDEX*. You and your library clientele will be glad you did.

isi®
Institute for Scientific Information
325 Chestnut Street

Publications of the
AMERICAN DOCUMENTATION INSTITUTE

AMERICAN DOCUMENTATION,* quarterly journal of the ADI. Subscription rate: $18.50 per year, plus $.50 postage for foreign subscriptions. (Subscription included in annual membership dues of $20.00.)

EDUCATION FOR INFORMATION SCIENCE,† proceedings of a symposium held September 1965. $2.00 to ADI members, $6.00 to nonmembers.

Papers and Proceedings of ADI Annual Meetings—

AUTOMATION AND SCIENTIFIC COMMUNICATION‡ (1963). $9.50 to ADI members, $12.50 to nonmembers.

Beginning with the 1964 volume, the ADI Annual Proceedings are in a numbered series—

PARAMETERS OF INFORMATION SCIENCE† (Volume I, 1964). $7.85 to ADI members, $15.75 to nonmembers.

PROCEEDINGS OF THE 1965 FID CONGRESS† (Volume II, 1965). $9.30 to ADI members, $10.95 to nonmembers.

PROGRESS IN INFORMATION SCIENCE AND TECHNOLOGY§ (Volume III, 1966). $12.00 to ADI members, $16.00 to nonmembers.

AND TWO NEW PUBLICATION VENTURES OF ADI—

DOCUMENTATION ABSTRACTS,‖ a quarterly abstract journal designed to be a comprehensive source of information about the literature of documentation and related areas—initiated in 1966 and published jointly by the American Documentation Institute, the Chemical Literature Division of the American Chemical Society, and the Special Libraries Association. 1967 subscription: $15.00 to ADI members, $25.00 to nonmembers.

ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY,# a new series devoted to consolidating the latest developments of the growing field of information science and technology. The series will not merely reflect or cater to current interests; it will attempt, also, to broaden and deepen them. Volume I, 1966, is $12.50 to nonmembers of ADI, and $10.63 to ADI members *if members order from the American Documentation Institute.*

*Order from

AMERICAN DOCUMENTATION INSTITUTE
2000 P Street, Northwest
WASHINGTON, D. C. 20036

PAYMENT WITH YOUR ORDER IS REQUESTED

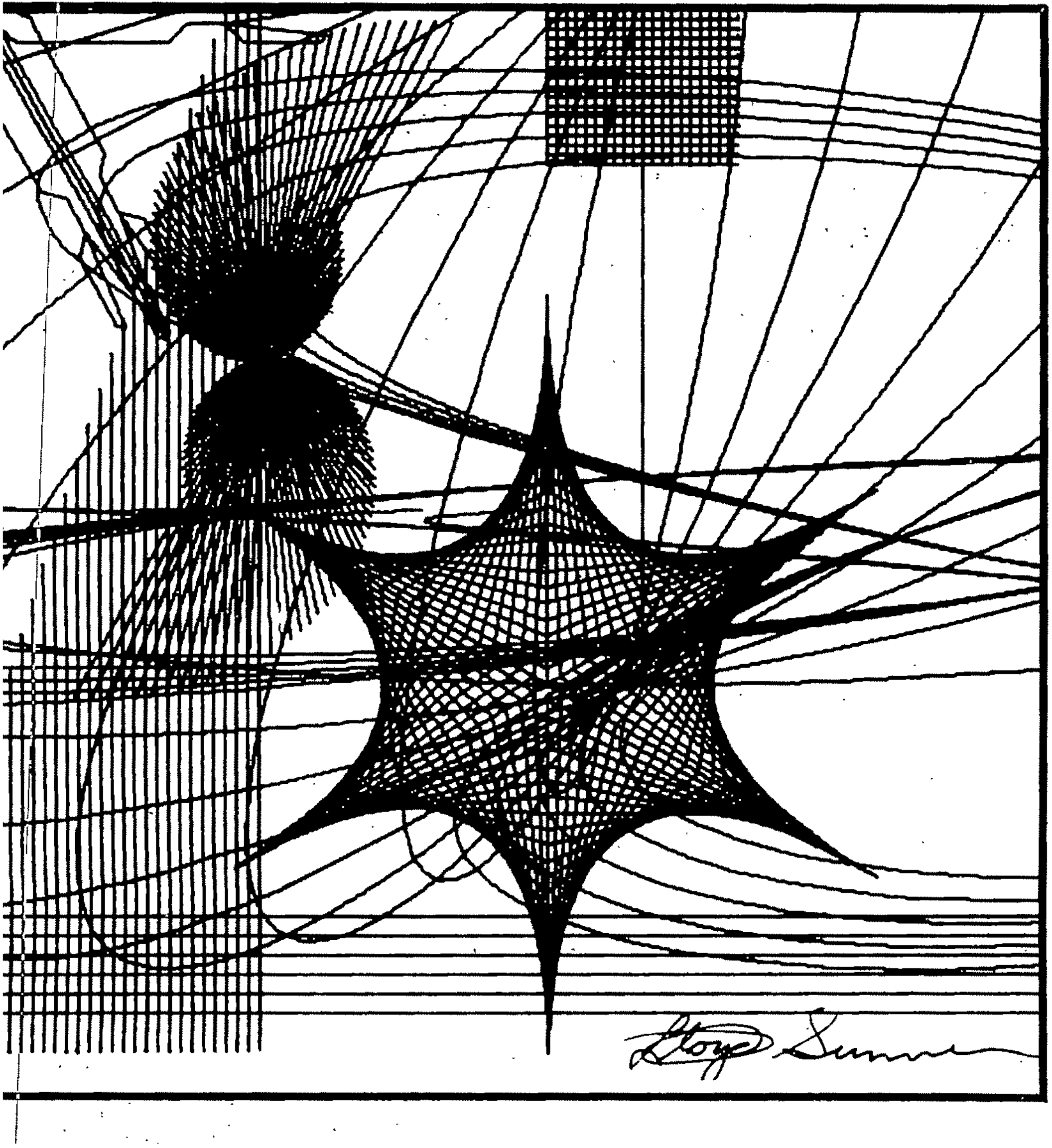
Other titles should be ordered as follows:

† Spartan Books, 1250 Connecticut Avenue, N.W., Washington, D. C.
‡ Kraus Reprint Corporation, 16 East 46th Street, New York, New York
§ Adrianne Press, P.O. Box 644, Woodland Hills, California
‖ Documentation Abstracts, Inc., P.O. Box 9018, Southeast Station, Washington, D. C. 20003
# John Wiley & Sons, 605 Third Avenue, New York, New York

merican Documentation

OCTOBER
1967
Vol. 18, No. 4

PUBLISHED QUARTERLY BY THE AMERICAN DOCUMENTATION INSTITUTE



30TH
ANNUAL
MEETING
AMERICAN
DOCUMENTATI
INSTITUTE
OCTOBER 23-2
NEW YORK HIL
NEW YORK
NEW YORK

AMERICAN DOCUMENTATION

INSTRUCTIONS TO AUTHORS

American Documentation is a publication of the American Documentation Institute. It is a scholarly journal in the various fields in documentation and serves as a forum for discussion and experimentation. Papers already published or in press elsewhere are not acceptable. For each proposed contribution, one original and two copies (in English only) should be mailed to Mr. Arthur W. Elias, Editor, *American Documentation,* Institute for Scientific Information, 325 Chestnut St., Philadelphia, Pennsylvania 19106. The manuscript should be mailed *flat* in a suitable-sized envelope. Graphic materials should be submitted with suitable cardboard backing.

Types of Manuscripts: Three types of contributions are considered for publication: full-length articles, brief communications of 1,000 words or less, and letters to the editor. Letters and brief communications can generally be published sooner than full-length manuscripts. Books, monographs, and reports are accepted for critical review. Two copies should be addressed to the Review Editor, Dr. T. Hines, 54 North Drive, East Brunswick, New Jersey.

Processing: Acknowledgment will be made of receipt of all manuscripts. *American Documentation* employs a reviewing procedure in which all mansucripts are sent to two referees for comment. When both referees have replied, copies of their comments are sent to authors with the Editor's decision as to acceptability. The refereeing procedure requires about 30 days. Authors receive galley proofs with a five-day allowance for corrections. Standard proofreading marks should be employed. Reprint order forms are forwarded with galleys.

Format: All contributions should be typewritten on white bond paper on one side only, leaving about 1.25 inches (or 3 cm) of space around all margins of standard, letter-size (8.5 × 11 inch) paper. Double spacing must be used throughout, including the title page, tables, legends, and references. The first page of the manuscript should carry both the first and last names of all authors, the institutions or organizations with which the authors are affiliated, and notation as to which author should receive the galleys for proofreading. All succeeding pages should carry the last name of the first author in the upper right-hand corner (0.5 inch from the top) and the number of the page.

Style: In general, style should follow the forms given in the Style Manual for Biological Journals (SMBJ), published for the Conference of Biological Editors by the American Institute of Biological Sciences (1964).

Title: The title should be as brief, specific, and descriptive as possible. Vague and unrevealing titles may delay publication.

Abstract: An informative abstract of 200 words or less must be included, typed with double spacing on a separate sheet. This abstract should present the scope of the work, methods, results, and conclusions.

Acknowledgments: Financial support may be listed as a footnote to the title. Credit for materials and technical assistance or advice may be cited in a section headed "Acknowledgments," which should appear at the end of the text. General use of footnotes in the text should be avoided.

Graphic materials: *American Documentation* requires finished artwork. Follow the style in current issues for layout and type faces in tables and figures. A table or figure should be constructed so as to be completely intelligible without further reference to the text. Lengthy tabulations of essentially similar data should be avoided.

Figures should be lettered in black India ink. Charts drawn in India ink should be so executed throughout, with no typewritten material included. Letters and numbers appearing in figures should be distinct and large enough so that no character will be less than 2 mm high after reduction. A line 0.4 mm wide reproduces satisfactorily when reduced by one-half. Graphs, charts, and photographs should be given consecutive figure numbers as they will appear in the text; however, figure numbers and legends should not appear as part of the figure, but should be typed double spaced on a separate sheet of paper. Each figure should be marked *lightly* on the back with the figure number, author's name, complete address, and shortened title of the paper.

For figures, the originals with two clearly legible reproductions (to be sent to referees) should accompany the manuscript. In the case of photographs, three glossy prints are required, preferably 8 × 10 inches.

Organization: In general, papers should state the background and purpose of the study, followed by details of methods, materials, procedures, and equipment. Findings, discussion, and conclusions should appear in that order. Appendixes may be employed where appropriate for extensive lists, statistics, and other supporting data.

Bibliography: Accuracy and adequacy of the references are the responsibility of the author. Therefore, literature cited should be checked carefully with the original publications. References to personal letters, abstracts of verbal reports, and other unedited material may be included. If an as-yet-unpublished paper would be helpful in the evaluation of a manuscript, it is advisable to make a copy of it available to the Editor. When a manuscript is one of a series of papers, the preceding member of the series should be included in literature cited.

Citation Format:

Order: Literature cited should be sequentially numbered as cited.

Authors: Give all authors with arrangement as follows:

Elias, A. W., B. H. Weil, and I. D. Welt

Titles: Give full titles of articles in English, indicating language of original as: (In Ger.)

Journals: Journal titles should be given in full.

Monograph and Serial Data: Should be presented in order as follows: Volume, issue number, pagination, and year. The issue number should be given in parentheses if journal pagination is not continuous from issue to issue. Pagination should be inclusive. Year of publication should be given in parentheses. An example is given below:

Bishop, D., A. L. Milner, and F. W. Roper, Publication Patterns of Scientific Serials, American Documentation, 16 (No. 2): 113–21 (1965).

American Documentation is published in January, April, July, and October. One copy is included in the individual membership fee ($20.00 per year), three copies in the contributing membership fee ($100.00 per year), and up to five copies in the sustaining membership fee ($500.00 per year). Nonmembers may subscribe at $18.50 per year, postpaid in the U.S. Single copies may be purchased for $4.65 each. Communications concerning memberships, subscriptions, reprints, renewals, back issues, advertising, and changes of address should be sent to the American Documentation Institute, 2000 P Street, NW, Washington, D. C. 20036.

American Documentation is indexed in *Library Literature, Current Contents of Space, Electronic & Physical Sciences, Library Science Abstracts, Science Citation Index, Chemical Abstracts,* and *Documentation Abstracts.*

American Documentation is entered for second class mailing at Baltimore, Maryland.

American Documentation

PUBLISHED QUARTERLY BY THE AMERICAN DOCUMENTATION INSTITUTE

Vol. 18, No. 4 OCTOBER 1967

COVER "Conspired Abstract of the Christmas Spirit" by Lloyd Sumner, created using Burroughs B 550 computer and Calcomp 565 digital plotter. Having been disturbed by the overcommercialization of what should be a religious holiday, Mr. Sumner programmed this composition in an effort to capture the unadulterated spirit of Christmas. We regret that space limitations enabled us to show only a part of this composition.

Published quarterly by AMERICAN DOCUMENTATION INSTITUTE and printed by Port City Press, Inc., 1323 Greenwood Road, Baltimore, Maryland 21208. Second class postage paid at Baltimore, Maryland.

A General Model of Information Transfer: Theme Paper 1968 Annual Convention

A general model of information transfer establishes a conceptual framework for contributed papers for the 1968 ADI Convention in Columbus, Ohio, October 20–24, 1968. The general model is an elaboration on the classic sender/channel/receiver model and presents a variety of alternative channels for information transfer including direct transfer, primary recorded media, archives, secondary recorded media, and information centers. Suggested areas for response in the form of contributed papers include costs, performance, benefits, functions, application of scientific and technical disciplines, research, vocabulary control, and language processing associated with information systems, science, and technology. A call for papers for the 1968 ADI Convention is included.

JOHN W. MURDOCK and
DAVID M. LISTON, JR.

Battelle Memorial Institute
Columbus Laboratories
Columbus, Ohio

• Prologue

Information Transfer!!—That's the theme of the 1968 Convention of the American Documentation Institute to be held in Columbus, Ohio, October 20–24, 1968.

The technical committee of the convention believes that most authors reporting on information work and research believe that their efforts will in some way improve the transfer of information. The committee plans to use this common interest to give a special coherence to both the convention and the published proceedings. The plan is to establish a conceptual structure for the technical program in the form of a general model presented in the following "theme paper" on information transfer. It is conceived that authors will be able to respond within their own specific areas to the broad structure established by the general model. To foster this process, the theme paper presents the general model and poses questions about many of the specific problem areas contained therein. Its purpose is to promote thought and response in the form of contributed papers which will provide the backbone of the technical program and be logically interrelated by the structure of the general model. Each contributing author will be requested to introduce his paper with a description of the correlation between the model and his specific subject area. Papers highly correlated with the theme will become the convention contributed papers. Other papers of high quality judged to be of interest to ADI members will provide the content for the author forums. Thus, the following theme paper heralds the call for papers for the 1968 ADI Convention. Specific suggestions appearing throughout the text for responding papers are set in italics to bring them obviously to the reader's attention.

• Introduction

Inherent in at least one set of definitions of the words "knowledge" and "information" is the concept that an item of knowledge becomes an item of information when it is "set in motion"—when it enters the active process of being communicated or transferred from one or more persons, groups, or organizations (sender) to one or more other persons, groups, or organizations (receiver). Many people will argue that knowledge as defined here has no intrinsic value—that only when it is successfully transferred is its value to be realized. Others go further, arguing that the value of information cannot be realized until it is actively applied in decision making. Either of these viewpoints must necessarily concede that *value* is dependent upon *transfer*. Thus, *information transfer* is an important and appropriate theme for the 1968 American Documentation Institute Convention. This theme paper presents a generalized model of information transfer to set the stage for the convention's technical program. The initial call for papers is included as the final section. *Some persons may wish to respond to the call for papers by exploring the idea of value being dependent on transfer.*

• The General Model

Figure 1 presents graphically the general model of information transfer. It is immediately obvious that the model is based on the classic sender channel receiver concept. In this case, there is a variety of alternative channels.

The Variety of Channels and the Communication Continuum

Communication between sender and receiver can occur at a number of levels along what is referred to as the "communication continuum." This also was called the "feedback dimension" by Lawrence Berul (*1*). The authors believe the general model in Fig. 1 includes every type of communication channel for information transfer. The value of the model is in the possible orientation or perspective that it provides for authors to say "Here is where my specialty helps in the information transfer." For example, in the situation of an individual who writes himself a note, the note is the primary recorded medium and his file of notes (or desk top or drawer) is the archive. He becomes the user when he wishes to retrieve the note. Sophistication is added when several people prepare reports or write memos and the archives become a central file. Further, complexity is added when the media include reports from outside the organization such as published literature. The archives now comprise a library or its equivalent. *It is possible similarly to relate other information work to the model.*

The Direct Channel. One extreme of the communication continuum (included in the direct, nonrecorded transfer channel of the model) is face-to-face discussion in which communication is:

1. Very direct.
2. Very dynamic, permitting the utilization of:
   - words, phrases, sentences, etc. (language);
   - gesticulations;
   - inflections of the voice;
   - interruptability, allowing the receiver to interrupt the sender requesting clarification of or elaboration on the message being spoken;
   - feedback, allowing the receiver to become the sender with reverse flow of information transfer;

3. Very rapid, with virtually no delay time involved. Disadvantages primarily relate to:

1. Faulty memory;
2. Little chance for study of what is transferred;
3. Frequent acceptability of vague generalizations which would not be permitted in a recorded message.

Progressing from the point of face-to-face discussion along the communication continuum toward situations involving less directness, less dynamic transfer, and more time delay, one can visualize situations such as phone conversations, television broadcasting, and radio broadcasting. All of these types of transfer are signified by the direct channel from the originator to user depicted in the general model.

The Primary Recorded Media Channel. Eventually the point is reached where the originator feels that what he has to say should be recorded as part of the body of literature of his discipline. This publication is usually thought of as the primary literature dealing with current topics. Until the past 5 years, little was done to package primary literature for retrospective searching other than providing periodic indexes. Probably much more could be done to make it readily retrievable. *It is hoped that someone will consider writing on this subject in response to this paper.* Other examples of primary recorded media are letters, newspapers, conference notes, technical reports, handbooks, monographs, texts, patents, and tapes. *Each of these media is worthy of papers on information transfer.*

The Archival Channel. Because the user is not always sensitized to the flow of messages through the more current channels, the archival channel has developed to store information for subsequent delayed usage when the user becomes aware of a need for it. Document depots, libraries, special libraries, and corporate files are all forms of archival storage. *Continued reporting of research on improvement of archives is hoped for as input to the 1968 Convention.*

The Secondary Recorded Media Channel. The next channel for the transfer of information involves the secondary sources or media. It feeds from both the primary recorded media and archives and also becomes archival when collected into libraries and other holdings. The purpose of the secondary recorded media channel is to assist people to search, more easily, an ever increasing volume of current and stored information for items of interest. Secondary media such as abstract journals, accessions bulletins, indexes, and bibliographies are faced with increasing volumes of literature and with pressure to reduce the time period for funnelling information from the other channels into the secondary media channel. This has increased costs sufficiently to make people question whether value received is worth the cost. *This controversy could lead to many interesting papers.*

The Information Center Channels. Information centers have increased in importance in the past 10 years. They represent an attempt to provide a service to essentially a known group of users upon demand. The information *analysis* center, in particular, attempts to utilize all information transfer channels to provide tech-

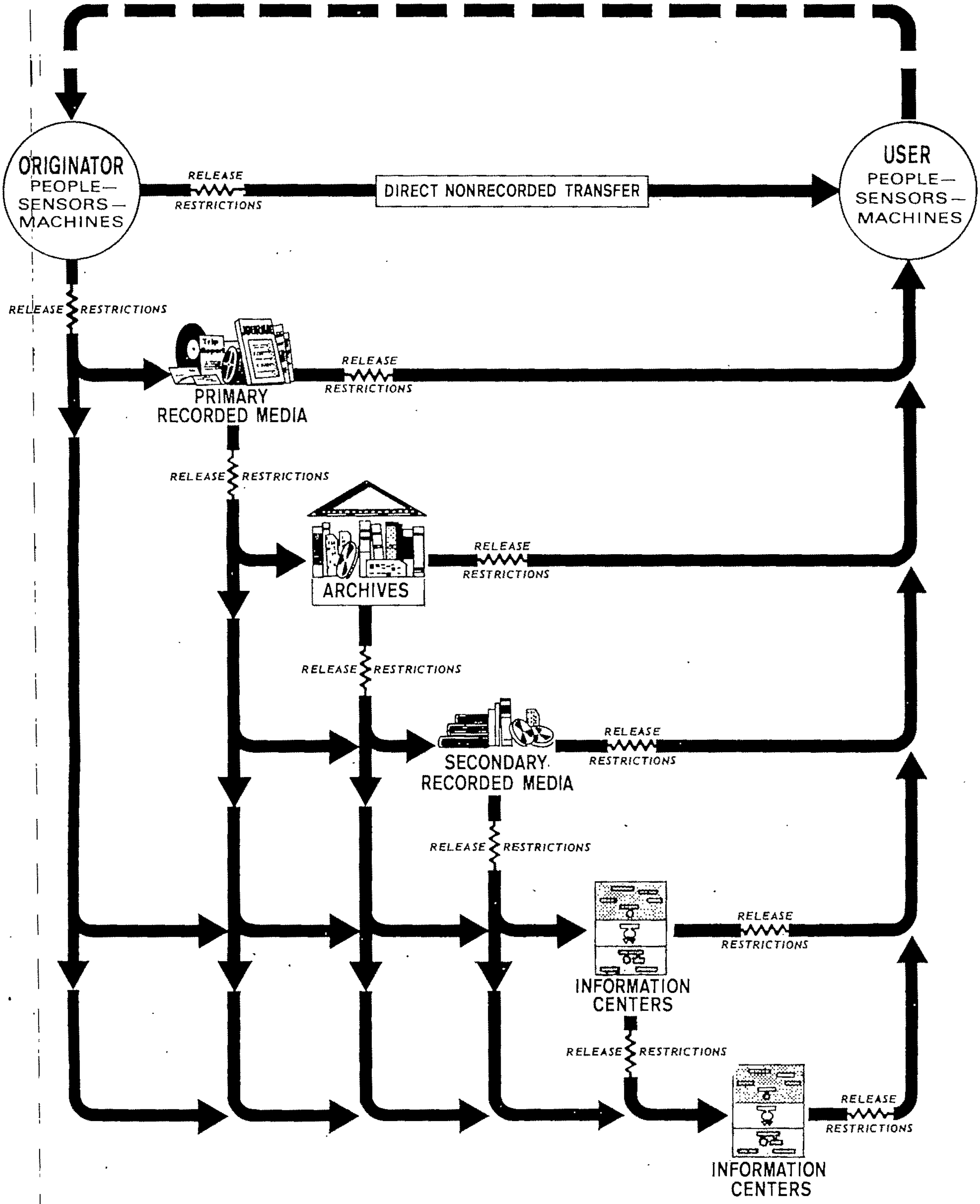


Fig. 1. General model of information transfer

nical answers to technical questions posed by users. Thinking in terms of an electrical analogy to the model, information centers act as "switching centers" utilizing the "circuitry" of the channels in the most appropriate combination of series and parallel arrangements.

The concept of analysis centers has been applied primarily to technical disciplines and mission-oriented projects. *Reports of applications of the analysis center concept to the social, political, and economic fields would be of interest for the 1968 Convention.* The functions and services of analysis centers were first described by G. S. Simpson (*2*) at the 1961 Boston ADI Annual Meeting. The symbol used in Fig. 1 to represent information centers was presented at that time and has been used in several conferences and papers since 1961. The three parallel segments of the symbol represent the primary functions of the analysis center as described by Simpson. The top segment represents the acquisition function; the middle segment represents the storage and retrieval function; and the bottom segment represents the primary function, analysis. In analysis centers as much as 80% of the budget is spent for the analysis of information by experts. The Special Interest Group of ADI on Analysis Centers is another recognition of the analysis center as an established activity in information transfer. Dr. Chalmers Sherwin said at the National Symposium on "Putting Information Retrieval to Work in the Office" on May 9, 1967, and in a paper (*3*) discussed at that meeting that he felt that the analysis center concept would provide the answer to the national information problem for at least the next generation. *This statement might prompt some responses which would be of interest at the 1968 Convention.*

The more often used expression, "information center," also has as its main characteristic the response to a customer on demand. However, the information center is distinguished from the information *analysis* center primarily by the lesser degree of analysis performed. Information centers respond to inquiries more specifically than libraries. For example, information centers often repackage information and often publish the new package. The primary functions of information centers are acquisition, storage/retrieval, and direct responses to customer's requests resulting in some publishing of special reports. Many hardware and system designers have worked on problems associated with improving information centers. *Papers on all facets of methods and mechanisms to improve the operations of information centers are encouraged by the Committee.*

Cyclic Nature of Transfer from Originator to User

In a gross sense at least, the entire information transfer model is cyclic in that users (as a group) are the same people, sensors, or machines as the originators (as a group). Even an individual has the problem of communicating with himself across the time span of present to future. This problem is especially important to the individual as he promises himself to return eventually to an item observed in the current literature which cannot be read currently for any of a number of reasons. In a generalized model of information systems, M. C. Yovits and R. L. Ernst of The Ohio State University also depict a cyclic flow. (*4*) In the Yovits/Ernst model (Fig. 2) the decision making function is analogous to the originator/user elements of the model in Fig. 1. The types of originators/users represented in the Fig. 1 model include:

- Individual people
- Individual sensors
- Individual machines
- Industrial corporations
- Not-for-profits
- Nonprofits
- Universities
- Professional societies
- Federal Government
- State Government

Release Restrictions

Regardless of the type of channel utilized in transferring a message, there are certain release restrictions which will impede the "free" transfer of information from originator to user. Returning to our electrical analogy, these release restrictions would be much like resistances or impedances in the circuits connecting the originators to the users. Furthermore, the total resistance to flow would probably vary according to whether the resistances in the channels were applied in series or in parallel or in combinations of both.

Everyone seems willing to grant that release restrictions are real phenomena. Even at the level of face-to-face communications, they exist in such forms as language difficulties, personal reluctances to divulge facts, and personal incapabilities of expression. Release restrictions become more noticeable as the contact between sender and receiver becomes progressively less direct—less face-to-face. One often does not write in a letter or say on the phone what he would say face-to-face. Thus, even though the release restrictions are not overtly applied, there is tacit adoption of restrictions as the contact between sender and receiver becomes more remote. But, there is much we do not understand about this impedance:

1. What is the magnitude of the impedance?
   What percentage of valuable information is not

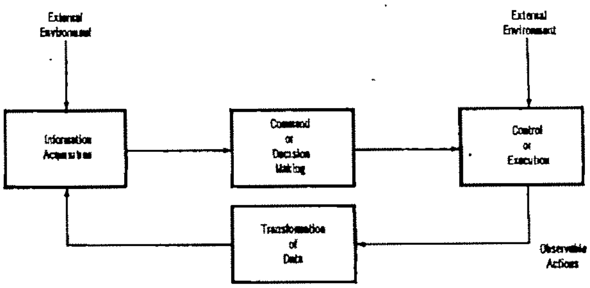


Fig. 2. Generalized information system model (*4*)

available to certain people because of security classifications, for example?

2. How critical is the impedance? To what extent does it really impair progress and understanding?
3. What possibilities are there for reducing or compensating for the impedance?
4. How justifiable are these impedances in view of the value of information—or do they exist *because of* the value of information?

Improved insight on these and related topics would be very worthwhile. Consideration might be given to the following different levels of restrictions (*5*):

(1) *Unclassified/Public Domain,* (2) *Unclassified/Copyrighted,* (3) *Personally Confidential,* (4) *Proprietary,* (5) *Security Classified,* (6) *Natural Language Discrepancy,* (7) *Personal limitations in written or verbal expression,* (8) *Expense* (*costs*).

• Some Specific Areas for Response

With the general model of information transfer serving as the underlying logical structure, a great many areas are made available for consideration. This section of the theme paper attempts to provide some preliminary inroads to some of these subject areas with the objective of promoting development of a full spectrum of papers on specific topics within this general framework. Such papers will be the heart of the technical program of the 1968 ADI Convention. The following discussions are not primarily to inform but rather to prompt thought and to invoke response. Authors may wish to discuss concepts that can apply at any point or combination of points in the transfer spectrum. A host of ideas for papers is inherent in our previous presentation of the general model of information. Some areas worthy of additional specific mention are:

- Cost/Performance/Benefit Interrelationships
- Functions Performed Within the Channels
- Scientific and Technical Disciplines Involved in Information Science and Technology
- Current Areas of Research
- Vocabulary Control/Language Processing
- Optimum Channel Utilization.

Cost/Performance/Benefit Interrelationships

As a sounding board for further discussion, we offer the following hypotheses concerning the interrelationships between costs, performance, and benefits of information systems. The term, costs, in this discussion simply refers to the costs involved in operating an information system. However, the clear definition of the other two terms is more critical to a clear understanding of the following discussion.

Performance. The term, performance, comprises the combination of five factors:

1. Coverage—the extent to which an information system covers all applicable information. There is inherently specified a theoretical finite portion of the total field of information which applies to the scope and mission of a system. Performance includes a measure of the completeness of coverage of that portion of information.

2. Usage—the extent to which the system serves all the information needs existing within its scope and mission. There is inherently defined a theoretical finite portion of the total need for information which is able to be satisfied by the system. Performance includes a measure of the completeness of satisfaction of that portion of the total information need.

3. Accuracy—the degree of perfection with which the system can fit applicable information to specific expressions of need. This factor involves the familiar measures of relevance and recall.

4. Speed—the speed with which the system can perform its functions.

5. Output Quality—quality of products and/or services offered to the system users.

Benefits. The term "benefits," is expressible in terms such as:

1. The extent to which all inadvertent duplication of effort can be prevented.
2. The extent to which the planning and decision-making functions of any organization can be improved.
3. The extent to which synthesis of new ideas can be fostered through the manipulation and observation of information contained in an information system.

From the above definitions we see that performance of a system is a function of factors internal to, or controllable by, the system. This is in contrast to the factors bearing on benefits. These are external to or beyond the control of the information system.

Cost-Performance Relationship. Figure 3 depicts the relationship between cost and performance. Our hypoth-

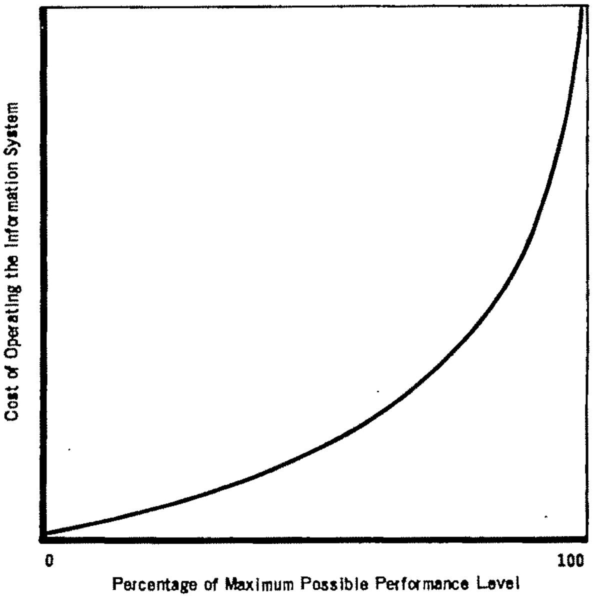


FIG. 3. Cost-performance relationship

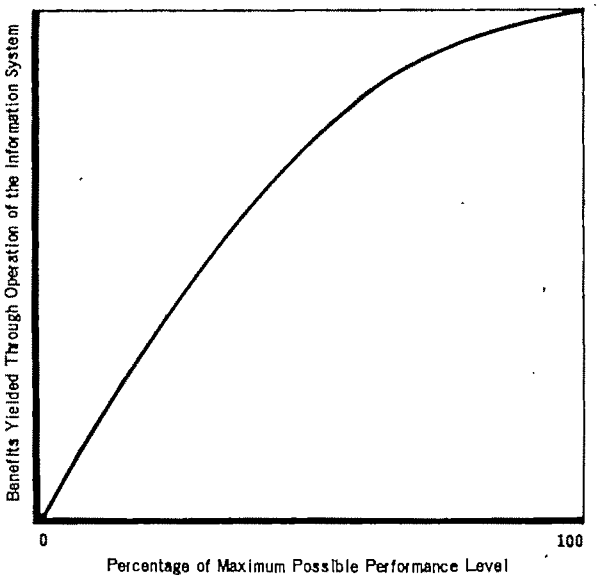


FIG. 4. Benefit-performance relationship

esis defines two basic characteristics of the interrelationship:

1. At zero performance level, the cost of operating the system is also zero.
2. As the performance level approaches 100%, the cost of operating the system approaches infinity.

Benefit-Performance Relationship. The relationship between benefits and performance is shown in Fig. 4. As the performance level increases, there will be a diminishing increment of benefit to be derived from each additional increment of performance—a tendency to approach a point of diminishing returns.

Benefit-Cost-Performance Relationship. Figure 5 presents the relationship between all three variables by plotting the benefit to cost ratio against performance. The benefit to cost ratio is similar in concept to return on investment. The shaded area represents conditions under which no information system should operate, because in this area it always costs more to operate the system than can be derived from it in the form of benefits. Curve C depicts an information system in a situation where there is no level of performance at which it can operate to produce a positive return on investment. Such a system would be completely unjustifiable. To operate optimally would be to operate at that level of performance (Point B) at which the system achieves the maximum benefit to cost ratio (Point X).

Cost-Performance Optimization. In Fig. 6, if Curve A represents the cost-performance relationship of an existing system, attempts at improving the system design toward optimum conditions (or designing the optimum system) can be represented as trying to "dent-in" the curve to arrive at a curve more like Curve B. This "denting-in" of the cost/performance curve can be accomplished by: 1. Devising ways to decrease costs without decreasing performance (as in moving from Point 1 to Point 2) in Fig. 6. 2. Devising ways to improve performance without increasing costs (as in moving from Point 1 to Point 3) 3. Devising improvements which combine Items (1) and (2), above (as in moving from Point 1 to Point 4).

Benefit-Cost-Performance Optimization. In Fig. 7, if Curve A represents the relationships for an existing system, then attempts at improving the system toward optimum conditions can be represented as trying to increase the value of the maximum benefit to cost ratio, regardless of the performance level at which the maximum ratio would occur. Examples of such improvements

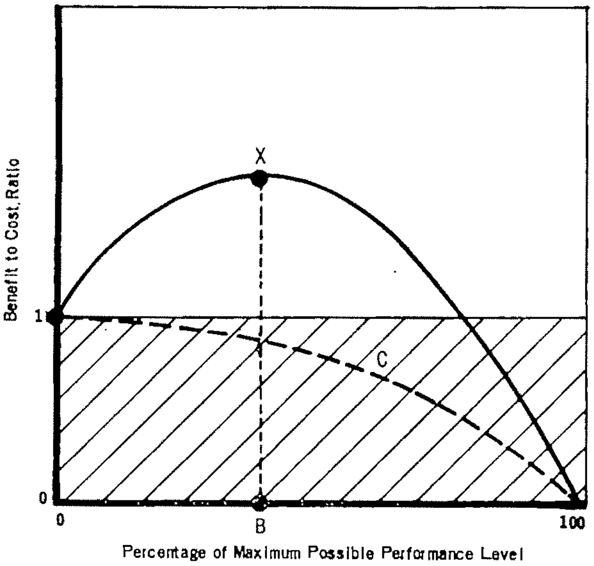


FIG. 5. Benefit-cost-performance relationship

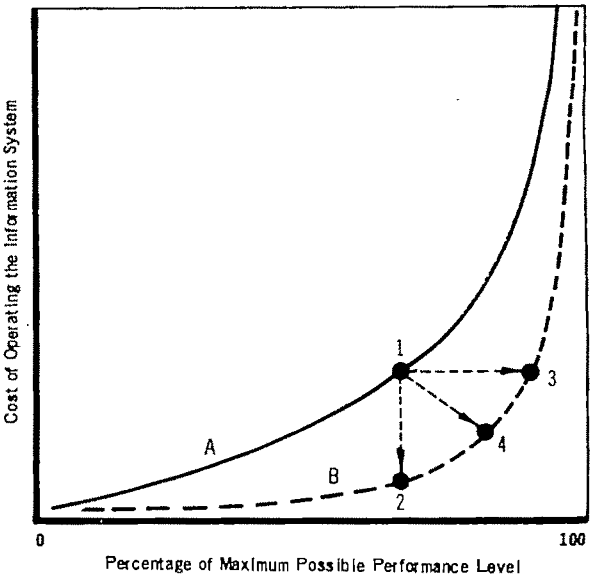


FIG. 6. Cost-performance optimization

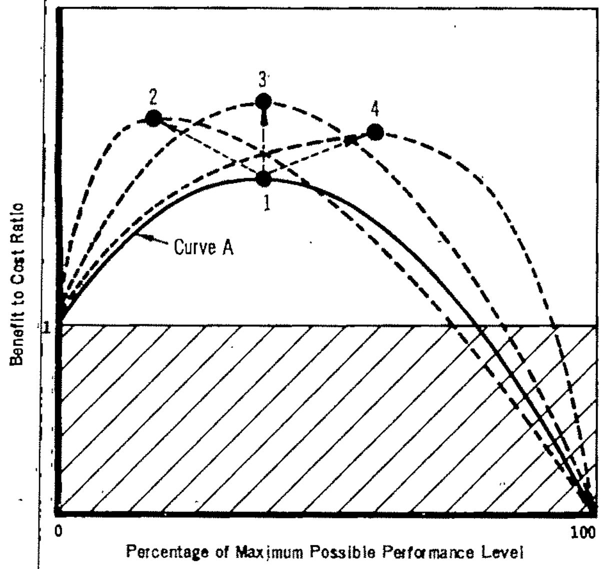


FIG. 7. Benefit-cost-performance optimization

are depicted in the figure as increasing the maximum ratio from Point 1 to any of the Points 2, 3, or 4.

The prime difficulty with making the above hypotheses a working tool is the elusive nature of the measurability of the factors involved. Take, for example, the cost factors. It seems a straightforward problem to measure costs of an information system. However, if the concept of "system" is extended (as it probably should be) to include the users and their costs of "doing business," the measurement of costs becomes very difficult. In that case the cost of *not* operating a system is *not* zero because the costs to the user of *not* having a system would have to be accounted for. The curve in Fig. 3 might, instead, be "U" shaped. Additional measurement problems include:

1. How do you measure the parameters of performance, coverage, usage, accuracy, speed, and quality of products?
2. How are benefits to be detected if they occur externally to the system?
3. How are benefits to be measured if they can be detected?

Papers on cost/effectiveness are heartily encouraged. Convincing management to spend increasing sums of money for information systems will become increasingly difficult without means for tangible dollar justification.

FUNCTIONS PERFORMED WITHIN THE CHANNELS

Within each channel, there is a variety of functions performed to make the channel operative. The comparison offered in Table 1 seems to indicate a fairly high degree of agreement between Wall (*6*), Simpson (*7*), and Berul (*1*) in identifying the nature of these functions. A much more generalized expression of functions was suggested by Ben-Ami Lipetz in a lecture before an ADI seminar in Columbus, Ohio, early in 1967. He offered the view that all of these functions can be categorized into three general types: (1) Matching of records, (2) Movement or physical displacement of records, (3) Creation of new records from old records. *All the aspects of system functions including those served*

TABLE 1. Functions performed within information transfer channels

Wall (*6*)	Simpson (*7*)	Berul (*1*)	Lipetz
Acquisition	Acquisition	(Origination) Acquisition	Physical comparison of records (matching)
Surrogation	Abstract preparation and dissemination	Surrogation • Cataloging • Abstracting • Indexing	
Announcement	Accession list preparation and dissemination	Announcement	Movement or physical displacement of records
Index operation	Index preparation and dissemination	Index operation	
Document management	Storage Retrieval Dissemination	Document management Retrieval Dissemination	
Correlation	Bibliography preparation Answering technical questions Analytical studies		Creation of new records from old records
Vocabulary control	Reference searching Referral services		

TABLE 2. Scientific and technical disciplines involved in information science and technology

Type of Endeavor	Disciplines										
	Mathematical Modeling	Simulation	Probability	Statistics	Linguistics	Computer Tech.	Systems Engin.	Operations Research	Economics	Value Engin.	Human Factors
1. Theory Development. This field involves efforts toward building theory under the wide variety of practices that have empirically developed as a result of the pressing necessity for operating information systems.	X	X	X	X	X						
2. System Design. Research in this field would be directed to making the design of information systems a systematic process.		X		X		X	X	X			
3. Human Replacement. The intellectual effort by humans continues to be the major cost factor in information transfer systems. This field encompasses all of the efforts to develop automatic techniques to replace human intellectual processes.	X	X	X	X	X	X		X			X
4. Language Accommodation. This field covers all sorts of techniques and devices required to accommodate the fact that languages are very inexact—and to make information transfer systems work in spite of that fact.			X	X	X	X	X				
5. System Operation. Research in this field would encompass all efforts toward improved efficiency and effectiveness of the operation of information systems.						X	X	X			
6. Philosophical Development. Efforts in this field would be directed specifically at the frontiers of information science—thought transmission, programmed learning, bionic applications. For example, an "information transfer philosopher" might ask such provocative questions as "Isn't it possible that the techniques of reading and writing are becoming obsolete as information transfer techniques?" Literally any and all disciplines will likely come into play in exploring the philosophical frontiers.	X	X	X	X	X	X	X	X	X	X	X
7. Economics. Emphasis in this field would be on the cost/benefit aspects of information transfer.							X	X	X	X	X
8. Language Redesign. This field is directed toward the evolution of an exact language to serve at least as the system language of information transfer systems and, perhaps, for extended use by authors and in other aspects of scientific communication.					X						
9. Human Factors. Information transfer systems will remain man-machine systems for many years to come despite efforts in Field (3) above. This field will encompass efforts to improve the understanding and efficiency of the human aspects of and contributions to information handling.	X		X	X							X

by hardware and software continue to provide important areas for research and development and thus, fruitful topics for discussion within the framework of information transfer for the 1968 ADI Convention.

Scientific and Technical Disciplines Involved in Information Science and Technology

The variety of efforts involved in research, development, technical services, consulting, and operations concerned with information transfer require inputs from a number of different scientific disciplines. As examples, nine areas of consideration serve to illustrate the diverse disciplinary contributions needed to attack the problems. Table 2 presents nine areas of endeavor and their associated disciplines. *Papers discussing any of the myriad aspects of the applications of scientific and technical disciplines to the problems of information transfer would be valuable contribtuions.*

Current Research

From many corners are heard comments deploring the wide gap between research efforts and practice in the field of information science and technology. Many people find it very difficult to foresee how the products of current research in the field will find their way into real-life applications of a practical nature. *For such people, any efforts to close the breach between researchers and practitioners would be welcome contributions.* Figure 8 presents an example of such an effort. It attempts to show how techniques such as:

- Automatic Abstracting
- Automatic Indexing
- Character (pattern) recognition
- Machine translation
- Automatic speech analysis

(presently in various phases of research) may fit into the fundamental document handling function of processing documents to produce indexes. Figure 8 also in-

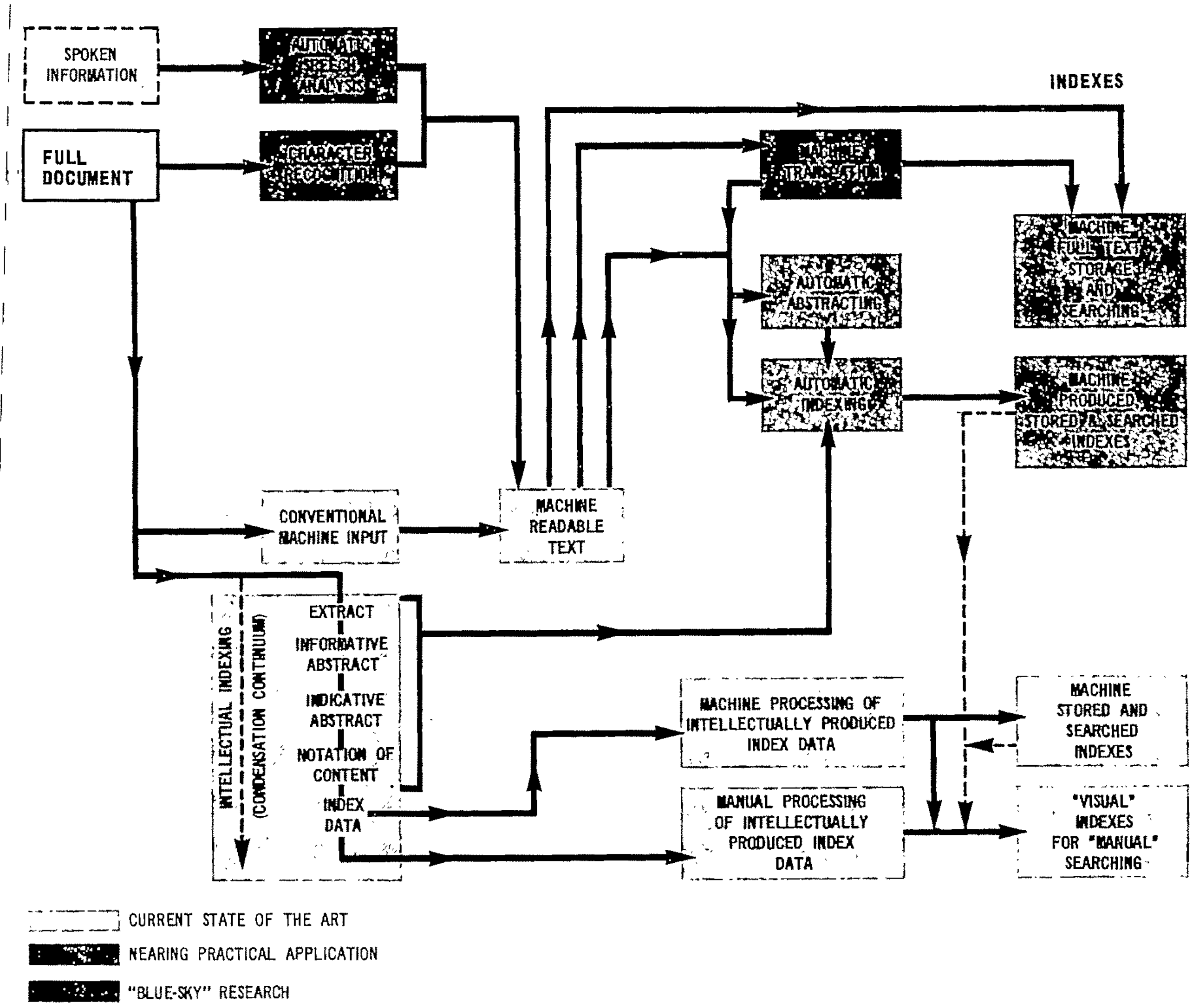

SPOKEN INFORMATION
AUTOMATIC SPEECH ANALYSIS
FULL DOCUMENT
CHARACTER RECOGNITION
INDEXES
MACHINE TRANSLATION
AUTOMATIC ABSTRACTING
AUTOMATIC INDEXING
MACHINE FULL TEXT STORAGE AND SEARCHING
MACHINE PRODUCED STORED & SEARCHED INDEXES
CONVENTIONAL MACHINE INPUT
MACHINE READABLE TEXT
INTELLECTUAL INDEXING
(CONDENSATION CONTINUUM)
EXTRACT
INFORMATIVE ABSTRACT
INDICATIVE ABSTRACT
NOTATION OF CONTENT
INDEX DATA
MACHINE PROCESSING OF INTELLECTUALLY PRODUCED INDEX DATA
MANUAL PROCESSING OF INTELLECTUALLY PRODUCED INDEX DATA
MACHINE STORED AND SEARCHED INDEXES
"VISUAL" INDEXES FOR "MANUAL" SEARCHING
CURRENT STATE OF THE ART
NEARING PRACTICAL APPLICATION
"BLUE-SKY" RESEARCH

Fig. 8. Current research applications

dicates those techniques which are now operational, those which are nearing practical application, and those which have more or less "blue-sky" status at the present time. Not covered by Fig. 8 are all of the various types of research methods which will be contributory to producing workable techniques of these types. *For the 1968 ADI Convention, papers on current research will be very much in order especially in two areas:* (1) *Papers presenting specific current research efforts;* (2) *Papers correlating such research efforts with eventual practical application as exemplified by Fig. 8.*

Vocabulary Control/Language Processing

All of the channels illustrated in our general model of information transfer are troubled with language difficulties. The language of the items of information entering any of the channels is not likely to provide a high level of similarity to the user's language to which the output from the channel must attempt to respond. Thus, there is usually a translation problem between input to and output from any of the channels.

At its worst, the translation problem will involve the conversion from one natural language to another. However, even when channel input and output are expressed in the same natural language, the correct matching of input and output ideas is plagued by a number of language problems:

- Semantics—the problem of word meanings including both synonyms (groups of words all having the same meaning) and homographs (single words each having more than one meaning)
- Generics—the problem of hierarchical word families
- Viewpoint—the problem of varying contexts as a result of varying viewpoints
- Term preconjunction—exemplified by the choice between the separate terms FLOW and RATE or the preconjoined term FLOW RATE as means for indexing a concept.

These language problems, it is claimed, produce adverse effects on the recall/relevance characteristics of information systems unless properly controlled. In many systems, the means of control has been the intellectually produced thesaurus. Rules for the intellectual construction of thesauri have been published by the Engineers Joint Council (*8*). Figure 9 depicts the parallel

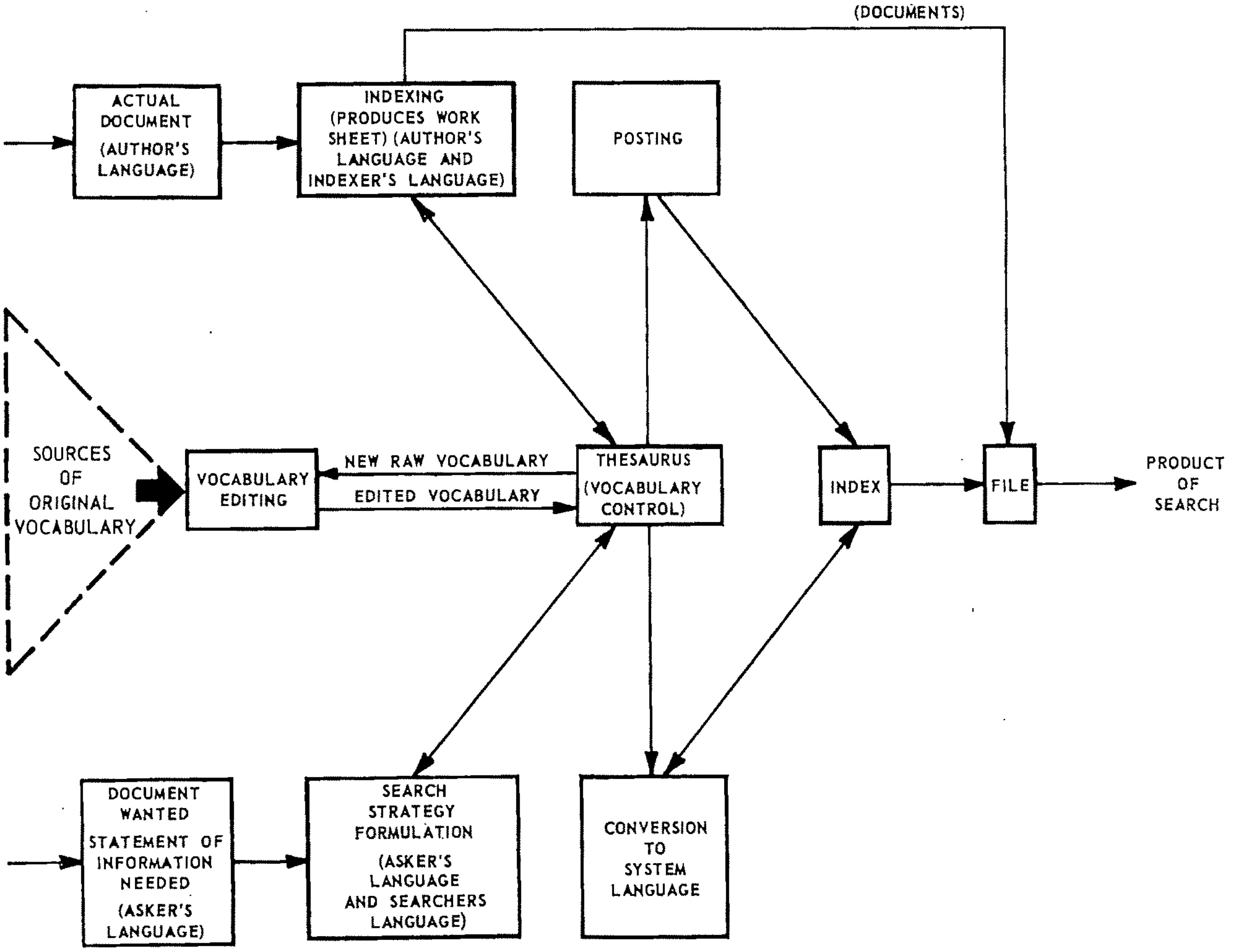

(DOCUMENTS)
ACTUAL DOCUMENT (AUTHOR'S LANGUAGE)
INDEXING (PRODUCES WORK SHEET) (AUTHOR'S LANGUAGE AND INDEXER'S LANGUAGE)
POSTING
SOURCES OF ORIGINAL VOCABULARY
VOCABULARY EDITING
NEW RAW VOCABULARY
EDITED VOCABULARY
THESAURUS (VOCABULARY CONTROL)
INDEX
FILE
PRODUCT OF SEARCH
DOCUMENT WANTED STATEMENT OF INFORMATION NEEDED (ASKER'S LANGUAGE)
SEARCH STRATEGY FORMULATION (ASKER'S LANGUAGE AND SEARCHERS LANGUAGE)
CONVERSION TO SYSTEM LANGUAGE

Fig. 9. The interactions between vocabulary control and the input/output elements of an information storage and retrieval system.

nature of the relationships between vocabulary control functions and the input and output functions of a typical information storage and retrieval system. In essence, the thesaurus creates a "system language" which is capable of translating or "understanding" both the language of the input items and the language of the users which is required for efficient output.

But, is the expensive process of intellectual thesaurus construction really necessary for obtaining good system performance? The second phase of the Cranfield Project (*9*) provides some evidence (and it is possible we may be oversimplifying our interpretation of the results) that the simpler the indexing language, the better the recall/relevance performance of a system as shown in Fig. 10. If the Cranfield results may be extrapolated to apply generally to all information systems, the need for elaborate thesauri may evaporate.

The term "language processing" seems to represent a much broader scope of consideration than the concept of vocabulary control discussed above. Robert F. Simmons (*10*), in the 1966 Annual Review of Information Science and Technology, organized his discussion of automated language processing as follows:

- Computational Linguistics
  - (1) Linguistic Theory
  - (2) Semantic Theory
  - (3) Psycholinguistics
  - (4) Automated Syntactic Analysis Systems
- Applications Studies
  - (1) Mechanical Translation
  - (2) Automated Question Answering
  - (3) Stylistic and Content Analysis

We feel that contributions in these areas and other areas dealing with language will provide many of the fundamental stepping stones to future improved methods for expressing ideas and concepts, for converting such expressions into storable/manipulable form, and for analyzing, and correlating elements of information and thus synthesizing them into new usable intelligence. These are functions which provide the underlying framework for improved information transfer.

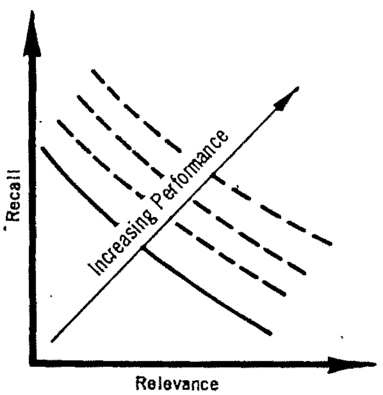


FIG. 10. Recall-relevance-performance characteristics

Optimal Channel Utilization

Figure 1 illustrates well that the user of information may have several options available to him when he has the need to obtain information. His choice may be limited by his resources or those of his organization. Often, however, the options are limited by the lack of awareness of the individual or his organization of the options available. There is also the possibility that the individual or his organization desires to improve the availability of information but hesitates to invest the capital into the development of this capability because of uncertainties in the value of the results to be obtained or in the choice of what system is best.

For example, most organizations when choosing to supply their members with assistance often establish libraries plus several services or specialized activities in addition to the library. Assume that, for dealing with published (or report) literature, an organization decides to provide additional specialized services to its members. The library, to meet this requirement, usually will procure hardware or services to deal with the published literature in an overall sense, such as classifying journals instead of articles in journals. To provide in-depth indexing, the library will likely increase its subscription to commercially available secondary journals and indexes. When a member of the organization develops a need beyond the commercially available services, then specialized storage and retrieval mechanisms are procured. In many cases, members of special programs and projects with extensive information requirements develop their own systems. In other cases, the management of the organization will authorize the development of large-scale mechanized information programs using computers, microimaging services, or other mechanisms. The multiple channels that may be used, and the variation of approaches within each channel, coupled with the inability to show (in quantative terms) return on investment, pose some interesting questions on optimizing channel utilization. The problem of choosing the optimum means within each channel is also a serious systems study. *The committee encourages the preparation of papers on the problem of optimal channel selection and the associated problems of choice within channels.*

• Epilogue—Call for Papers

This paper has set the theme and procedure of the 1968 ADI Annual Meeting in Columbus, Ohio, October 20–24, 1968. Those persons who intend to submit papers should notify David M. Liston, Jr., Battelle Memorial Institute, 505 King Avenue, Columbus, Ohio 43201 of their intent by March 1, 1968. It would be helpful if the subject of the intended paper could be given at this time and if possible the specific area of the general model of information transfer to which it will relate. A guide for authors will be sent to these persons immediately upon receipt of the notification of intent. Manuscripts must be received by D. M. Liston by May 1, 1968. Each person submitting a paper will be notified by July 1, 1968 whether his paper has been accepted.

References

1. BERUL, L., *Information Storage and Retrieval—A State of the Art Report,* Auerbach Corporation, Philadelphia, Pa., 1964.
2. SIMPSON, G. S., JR., Scientific Information Centers in the United States, *American Documentation,* 13 (No. 1): 1962.
3. SHERWIN, C. W., Evaluating and Compressing Scientific and Technical Information, *National Symposium on Putting Information Retrieval to Work in the Office* (1967).
4. YOVITS, M. C., and R. L. ERNST, Generalized Information Systems, The Ohio State University, paper presented at the Second Conference on Electronic Information Handling, Pittsburgh, Pennsylvania, 1967.
5. MURDOCK, J. W., and G. S. SIMPSON, JR., $'s and Secrets, *American Documentation,* 18 (No. 2): 110 (1967).
6. WALL, E., A Rationale for Attacking Information Problems, *American Documentation,* 18 (No. 2): 97–103 (1967).
7. SIMPSON, G. S., JR., and C. FLANAGAN, Information Centers and Services, *in* C. A. Cuadra, ed., *Annual Review of Information and Technology,* Interscience Publishers, New York, 1966, pp. 305–335.
8. SPEIGHT, F. Y., ed., *Guide for Source Indexing and Abstracting of the Engineering Literature,* Engineers Joint Council, New York, 1967.
9. CLEVERDON, C., and M. KEEN, *Factors Determining the Performance of Indexing Systems,* Vol. II, ASLIB Cranfield Research Project, Wharley End, Bedford, England, 1966.
10. SIMMONS, R. F., Automated Language Processing, in C. A. Cuadra, ed., *Annual Review of Information Science and Technology,* Interscience Publishers, New York, 1966, pp. 137–169.

Computer Usage in the Development of a Water Resources Thesaurus*

This paper describes one method by which a thesaurus has been developed making extensive use of a computer to supplement the intellectual effort. The computer techniques incorporate several exceptionally useful innovations not previously disclosed in the open literature. The thesaurus which was developed is similar in convention and format to that recently adopted for Project LEX (4).

DAVID F. HERSEY and WILLIAM HAMMOND †

• Introduction

Since the publication of the *ASTIA Thesaurus of Descriptors* in 1960 (*1*), the general trend among the large information facilities has been toward more rigid controls of their subject indexing vocabularies (*2*). This trend was given added emphasis with the publication of the Engineers Joint Council's *Thesaurus of Engineering Terms* (*3*) in 1964. The conventions and format of this latter publication have recently been adopted, with minor exceptions, by the Department of Defense (*4*).

In view of this widespread trend toward the employment of thesauri conforming to commonly accepted conventions and format, it is believed that the recent experience of the Science Information Exchange (SIE) of the Smithsonian Institution in developing such a thesaurus will be a timely contribution to the state-of-the-art.

This paper is in no way concerned with the thesaurus as a functional document, there being many advantages and disadvantages to this and other systems of indexing and retrieval. Its purpose is to review the recent development of a thesaurus for water resources research. Particular emphasis in this paper is given to the novel approach to thesaurus construction in which the intellectual contribution was manipulated by the computer to produce a thesaurus similar to EJC and LEX conventions and format.

The present work began in the fall of 1965 under a contract between the Science Information Exchange and the Office of Water Resources Research (OWRR) of the Department of the Interior, and resulted ultimately in a publication, *Water Resources Thesaurus* (*5*). The project was undertaken at the request of the OWRR to develop a word list which might prove useful to them and others working in the field of water resources research. Thus, it was essential that the selection and display of terms in a water resources thesaurus be compatible with the generally accepted usage of the research terminology in that field, as well as methods of indexing (*6*) (*7*) currently in use in the cataloging of water resources research.

• Methodology—General

The development of the thesaurus was the work of a large number of people. There were four general categories of participants: the SIE staff scientists; the lexicographic consultants from Battelle Memorial Institute; general scientific and information specialist-type consultants; and SIE and ARIES Corporation computer programmers and consultants who participated in the computer aspects of the project. The principal thrust in the present paper involves the detailed description of the effort of the latter participants. It is necessary to devote some space to a brief description of the intellectual effort which preceded the computer manipulations so that the actual role of computer can be more clearly visualized.

A preliminary vocabulary for the thesaurus was derived from the current SIE word lists (*6*) and from terms used in the indexing of projects in the *Water Resources Research Catalog* (*8*) (*9*). To these were added words obtained from government and nongovernment contributors who were asked to supply additional terms of value

* This work was supported by the Office of Water Resources Research, Department of the Interior, Contract #14-01-0001-720 and carried out under the senior author who served as project director.

† Respectively, Science Information Exchange, Smithsonian Institution, Washington, D.C. 20086; and ARIES Corporation, Westgate Research Park, McLean, Virginia 22101.

to individuals in their own fields of specialized interest, but using terms of general interest to workers in the broad area of water resources research. The real challenge here, of course, was to take water resources research, which contains a myriad of multidisciplinary aspects, and to develop a word list which would be helpful to users at various levels of sophistication. The final selection of main terms, and the determination of the "thesaurus relationships" among those terms, was made by a limited number of individuals who carried the burden of exercising the necessary intellectual judgment. These individuals, however, were completely conversant with the subject area—and the intended use—for which the thesaurus was being developed.

As each term was selected, a lead-term work card was prepared as shown in Fig. 1. Each lead term was then keypunched according to the instructions contained in Table 1 and the example illustrated in Fig. 2. Upon completion of this task, the punched cards were sorted alphabetically and printed out by computer. A seven-digit numeric lead-term code was assigned manually, after leaving a gap for insertion of additional terms. This code was then entered on the term work card and keypunched into the lead-term punched cards. The punched cards were then sorted into letter-by-letter sequence (utilizing the numeric sequencing codes) and a second print out was made. Subterms, together with their numeric sequencing codes, were then added to the term work cards. The subterms included other terms in the vocabulary determined to be broader than (BT), narrower than (NT), related to (RT), or use references (USE) for respective lead terms. The subterm entries were keypunched according to the instructions contained in Table 1, sorted together with the lead-term punched cards into letter-by-letter sequence and a preliminary edition of the thesaurus was printed out on the computer for review and edit.

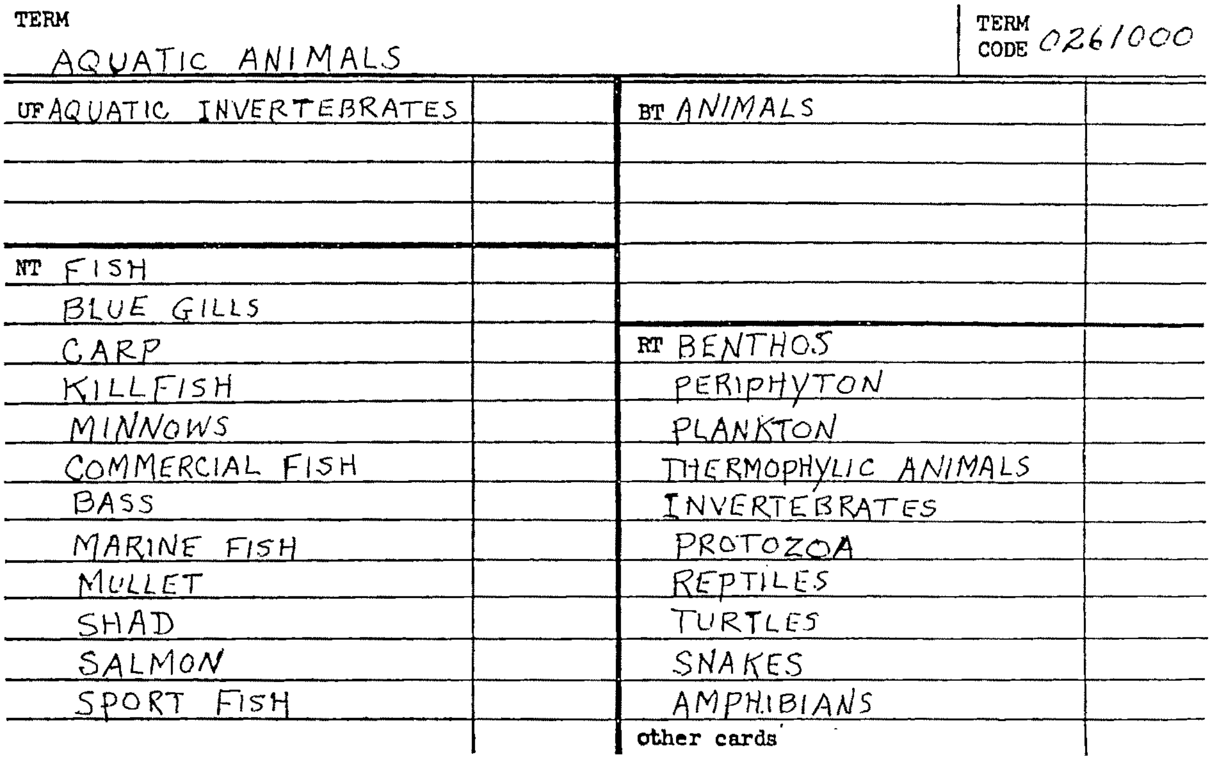
• Computer Role—General

The computer served a twofold purpose. It was used to maintain the integrity of the intellectual decisions reached concerning sequence coding, spelling, and immediate generic relationships among the terms. The computer was also used to generate the implicit generics among the terms throughout the thesaurus and to display these relationships for intellectual review.

One set of programs was written to edit the preliminary thesaurus compilation to insure that the initial corpus was in compliance with the given conventions and specifications. A second computer application was designed and a set of programs was written to provide a computerized capability for updating and maintaining the thesaurus on magnetic tape and for reproduction copy.

Essentially, the computer performed the following functions:

1. Edit for data format.
2. Edit for consistency in spelling and sequence coding.
3. Generate direct reciprocals for all subterm entries.



TERM AQUATIC ANIMALS

TERM CODE 0261000

UF AQUATIC INVERTEBRATES		BT ANIMALS	
NT FISH			
BLUE GILLS			
CARP		RT BENTHOS	
KILLFISH		PERIPHYTON	
MINNOWS		PLANKTON	
COMMERCIAL FISH		THERMOPHYLIC ANIMALS	
BASS		INVERTEBRATES	
MARINE FISH		PROTOZOA	
MULLET		REPTILES	
SHAD		TURTLES	
SALMON		SNAKES	
SPORT FISH		AMPHIBIANS	
		other cards	

FIG. 1. Lead term work card

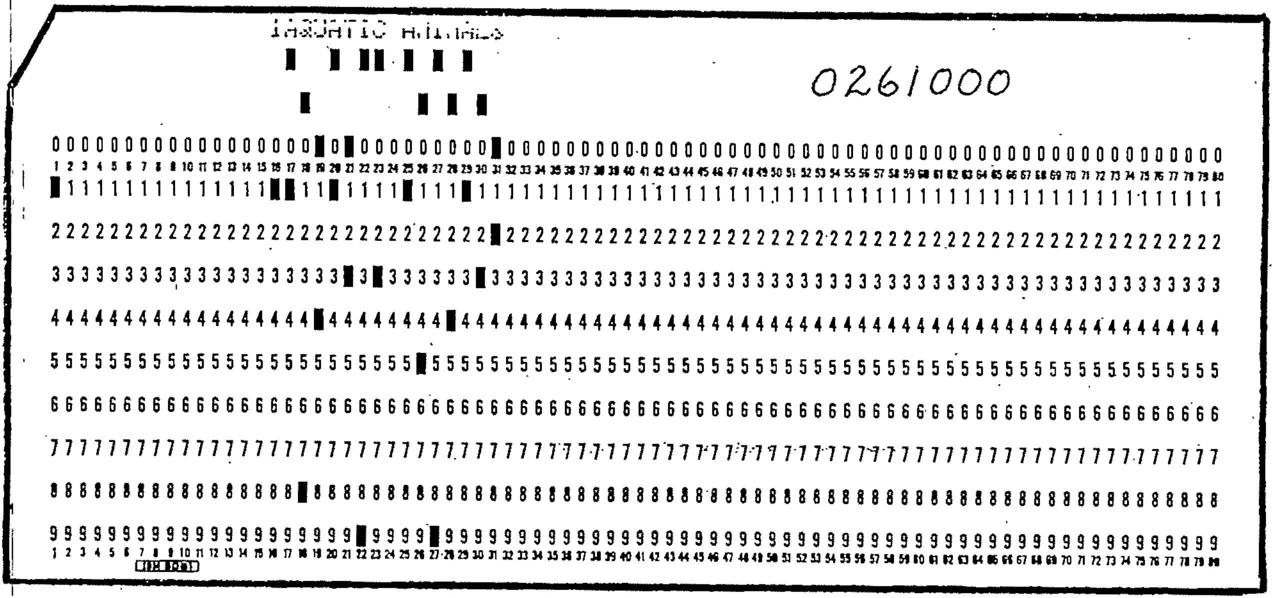


FIG. 2. Keypunched lead term

4. Generate "generic trees" for each of the main term entries—i.e., when a broad subterm (BT) is listed, all terms broader than it are also listed as BT subterms for the given main entry; when a narrower (NT) is listed, all terms narrower than it are also listed as NT subterms for the given main term entry.
5. Eliminate duplication and conflicting thesaurus relationships among the subterm set for a given main term entry.
6. Tag all terms in the file that are not of the lowest generic level, i.e., terms having narrower subterms listed in the thesaurus.
7. Generate major subject category groupings for the terms on the basis of the generic tree structures displayed in the thesaurus.

Once the initial thesaurus edit has been completed and all the conventions and specifications satisfied, new terms can be added or deleted, with assurance that the integrity of the "thesaurus relationships" among the terms in the vocabulary will be maintained.

Terms can be deleted on matching sequence code. The main entry record together with all its subterms, and all references to the term throughout the thesaurus can be deleted as a single maintenance action. A new term to be added must carry its immediate BT's *or* its immediate NT's and its RT's (immediate BT's are most desirable from the computer "logistics" viewpoint since more than one card would seldom be required to construct the generic tree for a new entry).

For the new term entry, the computer will automatically form the direct reciprocal and then generate the implicit generics (generic "trees") among the other terms in the thesaurus as well as the new terms being added. In the course of the maintenance run, the computer will eliminate conflicting thesaurus relationships among the subterm entries and will adjust the generic tag throughout the thesaurus to conform to any change in the generic status among the terms resulting from the update operation.

• Specific Application—Thesaurus Construction

The thesaurus construction (cleanup) programs were originally intended for one-time use. They were designed for processing only a single step at a time as work progressed on the initial compilation of the *Water Resources Thesaurus.* In the actual operation, however, a preliminary edition of the thesaurus, as discussed earlier, had been compiled and converted to punched cards before the computer application was written. For this reason, it appeared desirable to retain the existing punched card format for subsequent computer applications. This card format carries the terms and their numeric sequence codes, a numeric and alpha coding for the term relationship (BT, NT, etc.) and a numeric code for sequentially arranging successive scope note lines—all according to the instructions contained in Table 1.

The numeric sequence codes were intellectually derived to provide letter-by-letter sequencing except for preceding numerics, which were ignored. Gaps were retained in the numeric codes between succeeding terms to permit insertion of new terms.

In both the initial construction of the thesaurus corpus and in subsequent maintenance operations, step-by-step processing must be accomplished in precisely the sequence that is specified. Indicated corrections to the file from one step must be made before proceeding to the next step. Errors in data fields and inconsistencies in coding and spelling detected by the computer edit must be corrected before the missing reciprocals can be generated. The missing reciprocal entries detected by the computer

TABLE 1. Thesaurus card formats

1. *Lead Term* Card Punching	
Column 1	—Punch "1"
Columns 2-8	—Punch term code for lead term
Columns 9-15	—Skip (i.e., leave blank)
Column 16	—Punch "1"; if alphabetic term description exceeds Columns 17-80, punch a second card, duplicating in it Columns 1-15, punch "2" in Column 16, and continue punching alphabetic term description in Columns 17-80 of second card
Columns 17-80	—Punch alphabetic term description
2. *Explanatory Note for Lead Term* Card Punching	
Column 1	—Punch "2"
Columns 2-8	—Punch term code for lead term
Columns 9-15	—Skip
Column 16	—Punch "1"; if alphabetic description for explanatory note for lead term exceeds Columns 19-80, etc., follow Column 16 instruction for *Lead Term* card punching
Columns 17-18	—Skip
Columns 19-80	—Punch alphabetic description for explanatory note for lead term
3. *"Use" Reference* Card Punching	
Column 1	—Punch "3"
Columns 2-8	—Punch term code for lead term
Columns 9-15	—Punch term code for the "use" reference
Column 16	—Punch "1"; if alphabetic description for "use" reference exceeds Columns 23-80, punch a second card, duplicating in it Columns 1-15, punch "2" in Column 16, and continue punching in it alphabetic description of "use" reference in Columns 23-80 of second card
Columns 17-18	—Skip
Columns 19-21	—Punch in word "use"
Column 22	—Skip
Column 23-80	—Punch alphabetic description for "use" reference
4. *"Used for" (UF) Reference* Card Punching	
5. *"Narrower Term" (NT)* Card Punching	
6. *"Broader Term" (BT)* Card Punching	
7. *"Related Term" (RT)* Card Punching	
Column 1	—Punch "4" for "used for" (UF) references —Punch "5" for "narrower terms" (NT) —Punch "6" for "broader terms" (BT) —Punch "7" for "related terms" (RT)
Columns 2-8	—Punch term code for lead term
Columns 9-15	—Punch UF, NT, BT, or RT term code, as indicated by punch in Column 1
Column 16	—Follow sequencing punching pattern explained for preceding instructions if UF, NT, BT, or RT alphabetic description exceeds Columns 22-80
Column 17-18	—Skip
Columns 19-20	—Punch UF, NT, BT, or RT as indicated by punch in Column 1
Column 21	—Skip
Columns 22-80	—Punch alphabetic description for UF, NT, BT, or RT reference, as indicated

must be added to the thesaurus corpus and all conflicts in subterm relationships must be resolved before the generic expansion for the BT-NT subterm entries can be made. By conflicts of this sort, we refer to a given subentry having more than one relation (BT, NT, RT, USE, UF) to its corresponding lead-term entry—or added subterms following a USE entry. The generic expansion program will generate the implicit hierarchical structure among the BT-NT entries and will assure that a complete hierarchical BT-NT list is displayed under each lead entry in the thesaurus.

When all the above steps have been completed and corrective actions taken, the computer can tag terms that are not of the lowest generic level whenever they are displayed in the thesaurus.

• Thesaurus Maintenance

As originally conceived, the maintenance application would permit sufficient updating of a 10% or smaller increment to the thesaurus. In the actual processing, the programs were utilized for rather sizeable additions and numerous shiftings of term relationships. In one instance, this resulted in a 100% expansion of the thesaurus corpus. The program design specifications were sufficiently flexible to handle this type of operation, but when used in this manner were little, if any, more efficient than the equivalent cleanup programs. As a general rule, in an operational environment thesaurus updating and republishing would be infrequent. Therefore, simplicity of application was emphasized rather than computer efficiency.

The punch card input formats utilized in the thesaurus construction application were retained for maintenance, except that one field was added to record the maintenance operation desired. Table 2 contains instructions for preparation of the maintenance input.

In actual practice, maintenance was made unduly cumbersome because of the manual sequence coding. Conflicts between coding and existing sequencing, particularly for deletion and addition to the same main entry, could not be resolved by the computer without excessive programming effort. The immediate solution—again governed by the concept of simplicity of application—was to run deletions prior to processing any additions to the file. A single item could be deleted (and replaced by a correction) when additions were made to the file; however, deletion of any lead term and all of its subterm occurrences in the file (an "all-points" delete) had to be run prior to processing any additions.

TABLE 2. Thesaurus maintenance operation coding

	CARD COLUMN					
MOP*	1	77	78	79	80	OPERATION
1	1–7	0	–	–	1	Delete single line entry or entire scope note
2	1	1	0	–	1	Add new lead term without scope note
3	1	1	N†	–	1	Add new lead term with scope note of N cards
4	2	1	N†	–	1	Add scope note of N cards
5	3	1	–	0	1	Add new USE term and form reciprocal ‡
6	3	1	–	1	1	Add new USE term but do not form reciprocal
7	4	1	–	0	1	Add new UF term and form reciprocal ‡
8	4	1	–	1	1	Add new UF term but do not form reciprocal
9	5	1	–	0	1	Add new NT term and form reciprocal ‡
10	5	1	–	1	1	Add new NT term but do not form reciprocal
11	5	1	–	0	0	Add new NT term, form reciprocal,‡ and expand §
12	6	1	–	0	1	Add new BT term and form reciprocal ‡
13	6	1	–	1	1	Add new BT term but do not form reciprocal
14	6	1	–	0	0	Add new BT term, form reciprocal,‡ and expand §
15	7	1	–	0	1	Add new RT term and form reciprocal ‡
16	7	1	–	1	1	Add new RT term but do not form reciprocal
17	1	2	–	–	1	Delete all occurrences of term from file

* Maintenance Operation.
† No match between record types 1 and 2 is an error in Maintenance Run 3 (See Attachment 8).
‡ Reciprocals are formed in Maintenance Run 1.
§ Expansion is performed in Maintenance Run 5.
Any combination not found in the above table should be considered an error.

In the interest of simplicity, many modifications were made to the programs while the actual work was in progress. As an example, when the decision was made by the Office of Water Resources Research to include reciprocals for all related terms as well as for the remaining subterms, all of the missing reciprocals could be generated automatically by the computer and could be reentered without further review. In the earlier application an intellectual review was required to select missing reciprocals, which then had to be keypunched and added to the file.

Combined Application for Cleanup and Update Procedures

In further evaluation of the logic of the different computer runs, it was apparent that a combination of operational steps utilizing the best features from cleanup and maintenance would provide a simpler and far more efficient application for either—or both—functions. Deletions, however, still require special—and efficient—handling since two passes of the file are required to delete all cross references to a given lead term. Since updating the thesaurus should be infrequent, it was determined to be more economical—and far simpler—to accept this inefficiency rather than write a separate delete program.

An outline of the combined cleanup-maintenance application finally adopted is listed on Table 3.

When the combined thesaurus construction and maintenance application is employed for either purpose, the initial input (or update) corpus is keypunched in accordance with the instructions contained in Table 1 and with maintenance operation codes as shown in Table 2. The input data should contain at least:

Lead term entries
Scope notes
Use entries
Broad term entries—only the next broadest term is required to "thread" the BT-NT expansion. However, if more than one generic tree is involved, then the next broadest generic level for each tree must be entered.
Related term entries—no reciprocal required

Given the input data described above, the combined application shown in Table 3 will perform the edit functions, generate reciprocal UF's (use for) for each USE entry, generate a reciprocal RT entry for each given RT, generate an NT entry for each given BT, eliminate duplicate entries, eliminate—or tag—conflicts among the term relationships within the subterms of a given lead-term entry,[1] and then expand the BT and NT entries to form the full BT-NT generic set.

Inasmuch as all terms are under program control, those with narrower terms will be identified and will carry a distinguishing tag in the published thesaurus. The reproduction copy for final publication of the *Water Resources Thesaurus* was produced in single-column format.

[1] In this instance, the rules furnished through Office of Water Resources Research stipulated that: (1) if a main term also appears as a subterm, eliminate the subterm; (2) if the same term is a multiple entry within or among BT's, NT's, or RT's, first drop the direct duplicate appearing within same relationship category, then drop RT's and NT's (in that order) until only a single entry of the term remains.

TABLE 3. Combined cleanup-update application

STEP	PROGRAM	OPERATION
1	Cleanup #1	Card-to-tape and edit of existing Thesaurus corpus (or new additions to the Thesaurus).
2	Cleanup #2	Sorts output from Step 1 together with the existing Thesaurus master in a maintenance run as shown on Tape Record Ic below.
3	Cleanup #4	Sorts same file as in Step 2 to produce record as shown on Tape 1d below.
4	Cleanup #5	Punches out missing reciprocals from mismatch on output from Steps 2 and 4.
5	Cleanup #5A	Card output from Step 4 to tape.
6	Maintenance #4	Sorts output from Steps 2 and 5.
7	Maintenance #5	Expand full generic structure for BT-NT entries.
8	Cleanup #12	Sorts output from Steps 6 and 7.
9	Cleanup #13	Eliminates duplication and conflicts in thesaurus relationships among subterms.
10	Cleanup #10	Sorts output from Step 9.
11	Cleanup #11	Tags all occurrences of terms that have narrower terms listed in Thesaurus.
12	Conversion #1	Sorts output from Step 11 to Thesaurus master format for printing.
13	Thesaurus Print	Prints Thesaurus copy in single column continuous print.

Magnetic Tape Record
(80 characters, 10 records to a block)

			1	1	
1(N)	7(N)	7(N)	(O-N)	(1-7)	45 MAX (A-N)
A	B	C	C′	C″	D

Field	Size & Kind	Name
A	Fixed length field; 1 position; numeric	Type of record.
B	Fixed length field; 7 positions; numeric	Code of term in field.
C	Fixed length field; 7 positions; numeric	
C′	Fixed length field; 1 position; blank or numeric	Trailer card control.
C″	Fixed length field; 1 position; numeric	Reciprocal code for type of record.
D	Variable length field; 45 positions maximum	Actual alphanumeric as it appears in thesaurus.

Note: Character positions 63–79 reserved;
character position contains record mark.

Tape Ic

A	B	C	C′	C″	D
1	2183200	2183200		1	CEREAL CROPS
6	1623500	2183200		5	CEREAL CROPS
6	2274600	2183200		5	CEREAL CROPS
6	4625800	2183200		5	CEREAL CROPS
6	6113800	2183200		5	CEREAL CROPS
6	8491500	2183200		5	CEREAL CROPS
5	1310200	2183200		6	CEREAL CROPS
5	2348500	2183200		6	CEREAL CROPS
5	3872700	2183200		6	CEREAL CROPS
5	4136300	2183200		6	CEREAL CROPS
5	5234300	2183200		6	CEREAL CROPS

Tape Id

1	2183200	2183200		1	CEREAL CROPS
5	2183200	1623500		6	BARLEY
5	2183200	2274600		6	CORN, FIELD
5	2183200	4625800		6	OATS
5	2183200	6113800		6	RICE
5	2183200	8491500		6	WHEAT
6	2183200	1310200		5	AGRONOMIC CROPS
6	2183200	2348500		5	CROPS
6	2183200	3872700		5	GRASSES
6	2183200	4136300		5	MONOCOTS
6	2183200	5234300		5	PLANTS

Column and page make-up camera-ready copy was accomplished by the printer.

When the combined application is used for maintenance, deletions must still be processed first. The new entries can then be sorted together with the existing thesaurus master file in Steps 2 and 3 (Table 3). The rules for the new input remain the same as for the initial input described earlier. The processing also proceeds in a like manner; however, in step 7 an option is provided (Switch B is off) so that only the additional generic structure introduced by the new material will be generated, thus permitting faster update processing in steps 7, 8, and 9.

From actual operating experience, it was apparent that additional modifications to the programs could be made to reduce their running time. However, unless the simplicity of application would be greatly enhanced thereby to offset the added programming costs, these modifications were not made.

• Discussion

The development of a water resources thesaurus, as described in the present paper, attempted to combine the intellectual efforts of scientific specialists with the most advanced techniques in computer technology. A novel concept was introduced in the file organization for the computer application. Multilist, dual file records were employed, thus providing a practical magnetic tape application on the IBM 1460 (later run on IBM 360, M30). Understandably, the computer techniques developed originally, which seemed highly sophisticated at the time, have since given way to even more advanced techniques to better exploit the novel file organization first used in this project.

The publication of this original effort was intended primarily to disclose a technique and also to serve as a "stepping stone" or "jumping off" place to those who will be interested in further development of computer programs for vocabulary construction and control.

• Conclusion

This paper describes one method by which a thesaurus has been developed, making extensive use of a computer to supplement the intellectual effort. The computer techniques incorporate several exceptionally useful innovations not previously disclosed in the open literature on thesaurus development. The thesaurus developed in this study has been used in the indexing of a recent volume (*10*) of current *Water Resources Research Catalog* (in press).

Requests for further information on the applicability or availability of the computer programs may be addressed to the authors.

References

1. Armed Services Technical Information Agency, *Thesaurus of ASTIA Descriptors,* 1st ed., ASTIA, Washington, D.C., 1960.
2. Hammond, W., ARIES Corporation, Vocabulary Construction and Control, *in Proceedings of the Workshop on Working With Semi-automatic Documentation Systems,* (AD 620-360) Airlie Foundation, Warrenton, Va., May 2-5, 1965.
3. Engineers Joint Council, *Thesaurus of Engineering Terms,* 1st ed., Engineers Joint Council, New York, 1964.
4. Department of Defense, *Manual for Building a Technical Thesaurus,* (AD 632 279) U.S. Government Printing Office, Washington, D. C., 1966.
5. Office of Water Resources Research, U.S. Department of the Interior, *Water Resources Thesaurus,* U.S. Government Printing Office, 1966.
6. Unpublished Word Lists; currently in use at the Science Information Exchange, Smithsonian, Institution, Washington, D.C.
7. Kreysa, F. J., Science Information Exchange of the Smithsonian Institution, *in Proceedings of the Workshop on Working with Semi-automatic Documentation Systems,* (AD 620-360) Airlie Foundation, Warrenton, Va., 1965.
8. Office of Water Resources Research, U.S. Department of the Interior, *Water Resources Research Catalog,* vol. 1, pt. 1 and 2, U.S. Government Printing Office, Washington, D.C., 1965.
9. Office of Water Resources Research, U.S. Department of the Interior, *Water Resources Research Catalog,* Vol. 1, pt. 2, U.S. Government Printing Office, Washington, D.C., 1965.
10. Office of Water Resources, U.S. Department of the Interior, *Water Resources Research Catalog,* Vol. 2, U.S. Government Printing Office, Washington, D.C., 1966.

Analysis of Questions Addressed to a Medical Reference Retrieval System: Comparison of Question and System Terminologies*

Requests for subject and author searches submitted to the Medical Documentation Service of the Library of the College of Physicians of Philadelphia were studied. A total of 483 subject reference questions were analyzed for the number of question terms matching system subject headings (M), number of question terms translatable to system subject headings (T), number of stop-list words (S), and number of untranslatable words (U), using the judgment of the author. The average question had one M term or 22% M, one T term or 21% T, two S words or 38% S, and one U word or 19% U. Thus 81% of the average question was accounted for by M+T+S; in addition, 46% of the questions had no U words. Analysis of variance failed to show significant (5% level) differences between doctors' and lawyers' questions or between negotiated and nonnegotiated questions in number of M, T, S, or U. Study of limitations on searches for the 483 subject search requests and 38 requests for author searches showed that the requestor rarely stipulated type of material to be covered, languages to be included, time period to be covered, cost, or time for completion.

BARBARA FLOOD

School of Library Sciences
Drexel Institute of Technology
Philadelphia, Pennsylvania

• Definition of the Problem

One of the urgent problems confronting information scientists today is investigating the feasibility of direct interaction between user and system in various types of retrieval. This direct interaction is usually referred to as the man-system interface. There has been a great deal of research on the system side of the interface dealing with hardware, indexing, storage, and processing. There has been little investigation of the other side: How will the user approach the system?

The on-line use of computers for computational problems is experimental in many places; the computer's potential for augmenting human intellect in the computational sense is established; however, the potential with respect to reference retrieval is less clear. Such programs as Project MAC show that the problem of reference retrieval is under active investigation (*1*). Plans for automating the Library of Congress include the capability for subject access to the Library catalog via a query console (*2*). Project INTREX has similar goals (*3*).

We need to know something about how the user will approach the system with his reference question in order to plan for having the system responsive to the question. We need to know how the user will formulate this question. Most important, we need measures of the responsiveness of the system to the user's question.

Traditionally, the reference librarian mediates between the user and the system. The ideal reference librarian knows the user's needs and the system terminology, and

* Thesis submitted in partial fulfillment of the requirements for the Master of Science in Information Science, School of Library Science, Drexel Institute of Technology, January 1967.

can formulate the user's question in system terms. Can this mediation be conducted by a machine? There is a need to know the correlation between user and system terminologies.

Implicit in the question of the correlation between user and system terminologies is the problem of the terminologies of different user groups approaching the same system. In a system designed to serve a particular discipline (or mission), is the terminology of the user specialist significantly different from that of the nonspecialist? Presumably individuals state their questions in terms of their own disciplines. However, there is little information in support of this or about the terminology used by a nonspecialist in addressing a specialized system.

In addition to possible differences in terminology among different user groups approaching the same specialized system, there may be other interesting differences among reference questions. For example, are there differences in the limitations on the questions with respect to material to be covered, languages to be included, time period to be covered, etc? In other words, is it necessary to make different provision for different user groups approaching the same system?

The first purpose of this study was to develop a method for comparing user and system terminologies and to apply the method to a group of reference questions. The hypothesis tested was that different user groups would formulate their questions differently with respect to: (1) terminology and (2) limitations on searches.

The second purpose was to develop a method of characterizing the extent of mediation required between user and system, given different system components.

The specific objectives were then:

1. to develop a method of objectively characterizing questions and of comparing question and system terminologies;
2. to apply this method to a group of reference questions;
3. to analyze the differences among user groups and between negotiated and nonnegotiated questions with respect to:
   a. question terminology
   b. request limitations;
4. to analyze the extent of mediation that might be required between user and computer system for different systems.

• Material and Methods

Approach

A reference question can be classified in four ways: (1) by the characteristics of the questioner (user), (2) by the format of the request, (3) by limitations imposed on the request, and (4) by the terminology of the question itself. In the present study the user is characterized simply by the profession to which he belongs. Request format refers to what Berul (*4*) has called the feedback dimension in retrievability; that is, the immediacy of the interaction. The immediacy proceeds from person-to-person, to telephone, facsimile, letter, etc., to the remote interaction of addressing historical material. In the present study the format is classified as oral or written. The phrase "limitations on the request" includes anything the user might say about the request other than the question itself; examples are material to be searched, cost, and the time dimension of the literature to be covered.

There are other ways of classifying reference questions but most of them are derived from the answer to the question rather than from the question. Examples include various classifications of answers, sources used for searching, time taken in searching, and physical form of the answer (e.g., bibliographic list). These are not considered here.

The approach to the analysis of question terminology taken in this study includes certain background assumptions which require explanation.

For interaction with a user, a system has a certain set of components. There are different types and numbers of components in different systems. A reference retrieval system may have three typical components for interaction with the user. The first component is a list of system terms. A second is a list of rules for determining match to the list of system terms. A third is a list of words which will not appear in the system. That is to say, a list of entries, a list of translation algorithms, and a list of prevented words. Different systems will have these lists in different number and in different combinations.

It is possible to analyze questions addressed to a system in the context of the responsiveness of the different system components to question elements. This involves differentiating the question into analogous components and matching each question component to the corresponding system component; that is to say, classifying question elements in terms of system components. Such an analytic approach ignores the syntactical arrangement of the question components. In addition, the elements are not necessarily single words; each question element is a unit list match (unmatched elements are then words). The sum of the question elements which match the different lists and those which do not, is then an arbitrarily chosen quantity which represents the sum of the different kinds of matches and the unmatched words.

In the present study, the three system components are: (1) subject headings, (2) translations, and (3) nonsubstantive words. The four question components are (1) terms that are the same as the subject headings, (2) terms that can be translated to subject headings, (3) nonsubstantive words, and (4) words that do not fit into the other three categories. For example, in the request for material on "carcinoma metastasized to the esophagus," both "carcinoma" and "esophagus" are terms that are the same as subject headings. "Metastasized" can be translated to the subject heading "neoplasm metastasis." "To" and "the" are nonsubstantive words.

It is assumed that a reference retrieval system could contain any or all of the three lists. Both human and computer question analysis can proceed in terms of such lists. Measures of correspondance between question components and system components are therefore possible.

Source of Questions

The literature search records of the Medical Documentation Service (MDS) of the Library of the College of Physicians of Philadelphia were used as the source of questions. Although MDS deals solely in medical information, users include lawyers and other nonmedical persons as well as physicians. Physicians were therefore considered the "specialists" and lawyers and others were considered the "nonspecialists."

MDS charges a fee for its services. Therefore MDS deals with a particular class of information need: information the user is willing to pay for. It is assumed that such a need is close to at least one kind of need someone might have in approaching a remote terminal.

One source of questions to this service is the reference desk of the Library of the College of Physicians of Philadelphia. Questions are referred from the reference desk when the librarian estimates that the question will require more than 20 minutes of searching time to answer. (The reference desk service is nonfee.)

A second source of questions is direct communications from users. Direct communications may be by telephone, letter, or personal visit. This direct approach implies that the user is familiar with the service and has reason to believe that the service can handle his question.

A third source of questions is referral from a publishing company. The majority of the written requests originated in this way. The publishing company provides coupons for 1-hour searches with purchase of a medical encyclopedia. MDS has a contract with the publishing company to conduct these searches. The requests referred by the publishing company are slightly different from other MDS requests in two respects: (1) They are received indirectly and (2) they are not "fee" queries in the same sense as the others. However, the requestor had paid for the encyclopedia and it is assumed that the formalized act of specifying the request and sending in the coupon reflects a "need to know" similar to that reflected by contacting MDS directly.

In any case, the request is recorded on a search request form (Appendix A[1]). At the time of this study, records were available for 5 years, totaling 521 requests. For the most part, one individual[2] was responsible for recording questions and making searches during this 5-year period. Searches were occasionally delegated to others and requests were sometimes discussed with other staff members, but the same individual was responsible for all searches. For this reason, the problem of intersearcher reliability was not considered.

Questions from these three sources were divided into oral and written formats. Oral questions (telephone or in person requests) were probably "negotiated." That is, the searcher tried to resolve unclear points by asking the requestor what was meant. In this study such "negotiations" or changes of question terminology were taken as constant because they were conducted by the same searcher. The bias provided by the searcher's accumulating experience with the system and with the various types of questions over time is recognized but was assumed to apply to both "specialist" and "nonspecialist."

Negotiated questions, for this study, are those which were received orally, whereas nonnegotiated questions are those which were received in written form.

Material

The entire set of search requests received by MDS from October 1961 to May 1966 consisted of 521 questions; these were used for study. Because the system addressed was medical, the system language chosen was the list of Medical Subject Headings (MeSH) used by the National Library of Medicine for Medical Literature Analysis and Retrieval System (MEDLARS). MeSH was chosen because it is the main search tool at MDS.

Consideration was given to whether the edition of MeSH appropriate to the year of the question should be used. It was decided that a single edition (1965) would be used for all questions. (The problem of MeSH changes over time and how to standardize them is a separate one.)

Methods

A protocol sheet was filled out for each search request. The protocol sheet had the following information: (1) user profession, (2) format of the question, (3) search limitations, (4) verbatim question, and (5) question components (Appendix B). The questions were analyzed and tabulated and data concerning the question components were tested statistically.

User Profession

The user groups were "doctor," "lawyer," and "other." "Doctor" was defined as M.D., D.O., D.D.S., and the equivalent. This definition excluded paramedical personnel such as R.N., M.T., O.T.; these were included under "other." "Lawyer" included all requests from lawyers and law firms. "Other" included all requests not coming from "doctor" or "lawyer" as defined above.

Request Format

The request format could either be oral or written. Oral requests were those recorded by the searcher from

[1] Appendices on file at ADI Auxiliary Publications Service.

[2] Walter Bethel, Director, MDS.

an oral request received either on the telephone or in person. A written request was one received by letter, either directly or from the publishing company.

Search Limitations

Search limitations refers to comments about or restrictions on the request other than the question terminology itself. Limitations were tabulated as follows: *Type of material* included the choices: all, articles, and other. *Languages to be included* was grouped as English only, all, and other. *Time period to be covered* was grouped as current, 5 years, and other; time period means how far back in time to search the literature; for this study "current" was interpreted as coverage in the most recent year's literature. *Cost* was grouped as 1 hour ($8.), 1 hour preliminary (before authorization to proceed), and other; all the publishing company requests were automatically for 1 hour only. *Time for completion* was less than 1 week and more than 1 week; time referred to here is how soon the user wanted the material.

Terminology Analysis

Terminology analysis was conducted from the system side of the interface. That is, it was assumed that the system could contain various components, or lists. A match to an assumed list was considered one item regardless of how many words were in the item. Thus a subject heading item might be one word or it might be two or more words. Similarly a translation could be to a one word subject heading or to two or more words. In either case, the match was counted as unitary. Putting it another way, the match was to one list item rather than to one list word. The items on the assumed lists were those to which the question words were compared.

The analysis of terminology consisted of the author examining each question and finding out how many terms in the question were identical with terms in MeSH; how many terms could be accommodated by MeSH if they were translated into MeSH terminology; how many words could be taken care of if a stop-list (list of common words) were added to the system; and how many words were untranslatable. Thus there were four possible question components: (1) matching terms, (2) translatable terms, (3) stop-list words, and (4) untranslatable words.

Matching. There are two kinds of MeSH terms: subject headings and cross references. The two were recorded separately on the protocol sheet. In tabulation, however, there were too few cross references to warrant separate treatment.

Match was defined as an identical symbol string in the question and in MeSH. Note that a match may be one or several words long, depending on the MeSH entry. A list of matching terms was compiled and the frequency of each term was noted.

Translation. This question component resulted from human judgment about the nearest MeSH entry. The guide for translation was to be as objective as possible. There are varying degrees of subjectivity in translation as illustrated below.

Inversion occurs mainly with adjectival and prepositional phrases which are changed so that the noun is followed by its modifier in system entries.

e.g., Medical education = Education, medical
Dislocation of the hip = Hip dislocation

Word variants refer to words with the same root; i.e., plurals, verb, and adjectival forms.

e.g., Tumor = Tumors
Transplantation = Transplants

Compounding refers to a question term requiring translation to two or more MeSH entries or, conversely, two or more question terms comprising only one MeSH entry.

e.g., Deprol = Meprobamate + Benactyzine
Otomycosis = Mycosis + Otitis

Synonymy between question and MeSH terminologies could be difficult to evaluate as illustrated by the following example:

e.g., Pitfalls and complications of gallbladder surgery =
Postoperative complications + Cholecystectomy

The translation of "gallbladder surgery" to "cholecystectomy" might be considered a synonymous relation. Strictly speaking, however, the question term might refer to any operative procedure involving the gallbladder, not just removal (e.g., stone removal, duct stretching). The synonymy lies in the fact that cholecystectomy is the most common kind of gallbladder surgery and in the fact that MeSH carries no other entry which might comprise gallbladder surgery. This example illustrates the subjective nature of translation by an intermediary; the judgment depends on the intermediary's training and experience, both in the subject matter and with the system. To continue with this example, the other part of the translation ("postoperative complications") is even more questionable because it does not cover all "pitfalls and complications" which might occur before and during surgery. On the other hand, the alternative entry "surgery, operative," would appear too general to cover "pitfalls and complications."

Generic-specific relationships between question and MeSH terminologies comprised cases in which the MeSH entry was either more general or more specific than the question term. It was found early in the investigation that it was very difficult to find a cutoff point for generality or specificity. Therefore translations of this kind were held to a minimum.

A list of terms requiring translation was compiled and listed according to MeSH terms and also according to type of translation involved.

Stop-List. Another list which was compiled empirically was a list of common words considered to be nonsubstantive. Such words include articles, prepositions, adverbs, and conjunctions. The decision as to whether to put a word on the stop-list was made on the basis of (1) how "common" the word was judged to be and (2) how frequently it appeared. The stop-list was generated empirically so that it reflects judgment first and frequency second; in a further study, the basic list would be modified according to some frequency criterion. It was assumed that on-line machine searching could include a dictionary of nonsubstantive words similar to the list frequently used with KWIC programs. Each word for the stop-list was counted once (as opposed to each *term* for matching and translation).

Untranslatable Words. The remaining question component was the group of words which did not match, could not be translated, and were not stop-list words. These were called untranslatable words. A list of these words was compiled with notation of frequency. Examples are: recover, perforation, and secretion.

Summary. The question components included matches, translations, stop-list words, and untranslatable words. The sums of each can be formulated as follows:

$$Q = M + T + S + U \qquad (1)$$

where Q is the sum of the question components, M the sum of the matching terms, T the sum of the translatable terms, S the sum of the stop-list words, and U is the sum of the untranslatable words.

Examples of Question Analysis. Each question word, starting at the left, was looked up in MeSH, and each component was recorded.

Question 32. Carcinoma metastasized to the esophagus
M: carcinoma; esophagus
T: metastasized = neoplasm metastasis
S: to; the
U: —

Question 24. Isoenzymes of alkaline phosphatase
M: alkaline phosphatase
T: isoenzymes = enzymes
S: of
U: —

Question 21. Aneurysms of the uterine artery and its branches
M: —
T: aneurysms = aneurysm
uterine = uterus
artery = arteries
S: of; the; and; its
U: branches

Analyses

Analysis of Variance

The following hypotheses were tested for statistical significance at the 5% level using analysis of variance:

1. There is no difference in the number of matching terms among user groups or between request formats.
2. There is no difference in number of translatable terms among user groups or between request formats.
3. There is no difference in the number of stop-list words among user groups or between request formats.
4. There is no difference in number of untranslatable words among user groups or between request formats.

Measures of System Responsiveness

It has been assumed that a system has three possible components or lists to correspond to three terminologic components of the user's question. That is to say, there might be a component for matching terms (M), a component for translatable terms (T), and a component for stop-list words (S). By definition there is no system component for untranslatable words (U).

Given three components, there are seven possible system arrangements: (1) a match list alone, (2) translation list alone, (3) stop-list alone, (4) match and translatable lists, (5) match and stop-lists, (6) translatable and stop-lists, and (7) match and translatable and stop-lists.

We can therefore determine the ratio of matching terms (M), translatable terms (T), and stop-list words (S) in questions according to Equation 1, in order to be able to evaluate the responsiveness of different system arrangements to question terminology. The first thing we want to know is the equality (E) between question and system terminology; this is given by the match ratio expressed as a percentage:

$$E = M/Q \times 100 \qquad (2)$$

Then we want to know the improvement that would be effected by the various possible combined system arrangements. By adding matching and translatable terms we can derive a ratio which reflects the compatibility (C) between question and system terminology.

$$C = (M + T)/Q \times 100 \qquad (3)$$

This ratio reflects the proportion of the question terminology which is "substantive" from the system point of view. The other two two-way combinations may be calculated by addition; they are not considered of sufficient interest for separate treatment.

The three-way combination (match and translatable and stop) can be used to obtain the proportion of the question terminology that could be translated into system terminology. Translatability (Tr) is equal to the ratio of matching terms (M) and translatable terms (T) and stop-list words (S) to the total question components (Q).

$$Tr = (M + T + S)/Q \times 100 \qquad (4)$$

• Results

General

There was a total of 521 search requests. This total included 38 requests for author searches, 471 requests for subject searches, and 12 asking for both author and subject. The verbatim questions are listed in Appendix C. Only one of the author requests was for an historical name (Louis Pasteur, question 126).

User Profession and Request Format

The breakdown of user profession (doctor, lawyer, other) and request format (oral, written) is shown in Table 1. The "other" group included 36 requests from drug companies and scattered requests from students, laboratories, architects, librarians, hospital and surgical supply companies, and research foundations. No attempt was made to determine how many different individuals addressed the service. There were many examples of the same user asking different questions, both at the same time and at different times over the 5 years. There were also many examples of a user employing the service only once.

Search Limitations

Most requestors did not delimit the scope of the question. Examination of Table 2 shows that there were relatively few requests for all types of material, and when specified, articles tended to be asked for. In terms of languages to be included, there was a preponderance of requests for English language material only; the "other" group was small and scattered among European languages. Time period to be covered remained mostly unspecified. (The general policy at MDS is to search the most recent 5 years when time period is unspecified; author searches are generally for all publications.) The stipulations as to cost are skewed by the fact that all 142 publishing company requests were for 1 hour. Analysis of the time requested for completion showed that among those asking for the material in less than 1 week, 25 wanted it as soon as possible and 13 asked for same-day or 1-day service. In general, when one limitation was unspecified, so were all the others.

Table 1. User profession and request format

	Oral	Written		Total
		Publishing co.	Other	
Doctor	193			
	6*			
	199	137	15	351
Lawyer	80			
	32*			
	112	3	–	115
Other	31	2	22	55
Total	342	142 (179)	37	521

* Author.

Table 2. Search limitations

Type of Material	
Unspecified	413
All	26
Articles	66
Other	16
Languages to be Included	
Unspecified	384
English only	96
All	32
Other	9
Time Period to be Covered	
Unspecified	424
Current	30
Five years	22
Other	45
Cost	
Unspecified	328
One hour	155
One hour prelim.	0
Other	38
Time for Completion	
Unspecified	448
One week	65
More than one week	8

Terminology Analysis

The verbatim questions are given in Appendix C, the matching list in Appendix D, the translatable list in Appendix E, the stop list in Appendix G, and the untranslatable list in Appendix H.

The 483 subject questions generated 364 different matching terms which were used 559 times. The only matching terms that appeared frequently (8-10 times) were "cancer," "patients," "surgery," and "trauma."

There was a total of 397 different question terms that could be translated into 338 MeSH terms; these 397 terms occurred a total of 538 times. Analysis of the translatable terms is given in Appendix F in which the MeSH term is listed according to the type of translation involved. There were 21 examples of simple inversion. Word variants represented the largest group (138 times) and, of these, singular to plural or plural to singular translations occurred 67 times; synonyms occurred 42 times, including 5 examples of abbreviations; compounding occurred 20 times and generic-specific relationships 31 times. The "other" group (98 times) included cases of more than one type of translation as well as a few cases that did not fall in the above categories.

There were 98 words on the stop list, which occurred a total of 975 times. The frequencies of the 17 words occurring more than 10 times is given in Table 3.

TABLE 3. Most frequent stop-list words

Word	Frequency
of	227
the	114
in	113
and	98
to	36
or	29
on	25
for	23
as	20
with	19
a	19
by	17
effect	14
use	13
after	11
de	11
la	11

There were 362 different untranslatable words occurring 473 times. Examination of the list (Appendix H) shows that many of the words might be considered candidate stop-list words. There were also a number of substantive words such as "abasia," "arteriospasm," and "atresia," which were not considered translatable because they were too specific (i.e., the hierarchical distance from the nearest subject heading was too great).

The average question components and percentages are shown in Tables 4, 5, and 6. The average question contained about five terms, of which approximately one matched MeSH, one was translatable, two were stop-list words, and one was an untranslatable word. The range of Q was 1 to 25 units.

Table 7 gives a comparison of average number of question components and average number of question words in each category. The word data were obtained by counting the number and frequency of matching and translatable terms appearing in questions which had two or more words in these terms. The average question had six words in it.

TABLE 4. Average question components by profession

Profession	Q	M	T	S	U	N
Doctor						
oral	4.74	1.22	1.01	1.72	.81	193
written	5.65	1.13	1.27	2.26	.99	152
total	5.14	1.17	1.12	1.96	.89	345
Lawyer						
oral	4.81	1.00	.86	1.75	1.18	80
written	6.00	.33	2.00	2.00	1.67	3
total	4.85	.98	.91	1.78	1.19	83
Other						
oral	5.81	1.35	1.26	2.10	1.10	31
written	7.96	1.38	1.50	3.71	1.46	24
total	6.78	1.36	1.36	2.80	1.25	55
Total	5.28	1.16	1.11	2.02	.98	483

TABLE 5. Average question component percentages by format

Format	% M	% T	% S	% U	N
Oral					
Doctor	25.6	21.2	36.2	17.0	193
Lawyer	20.8	17.9	36.9	24.4	80
Other	23.3	21.7	36.1	18.9	31
Total oral	24.0	20.4	36.4	19.2	304
Written					
Doctor	19.9	22.5	39.9	17.6	152
Lawyer	5.6	33.3	33.3	27.8	3
Other	17.1	18.6	46.1	18.1	24
Total written	19.2	22.0	40.9	17.8	179
Grand Total	22.0	21.1	38.3	18.6	483

There were no appreciable differences among user groups in the percentages of M or S, but lawyers' questions tended to have a lower percentage of T and a higher percentage of U. Analysis by format showed that written requests tended to have a lower percentage of M, a higher percentage of T and S and a lower percentage of U as compared with oral. The differences in each case were small.

Analysis of variance failed to reveal significant differences among user professions or between formats in M, T, S, or U. Thus it was not possible to reject the null hypotheses. (Although the F levels required for 5% significance were 3.02 and 3.86, the only value above 1.00 was for between formats for T at 2.84.)

The effect of applying the various measures of system responsiveness (equality, compatibility, translatability) is shown in Tables 8 and 9. The remaining component is the untranslatable (*Untr*) fraction. The number of ques-

TABLE 6. Average question component percentages by profession

Profession	% M	% T	% S	% U	N
Doctor					
oral	25.6	21.2	36.2	17.0	193
written	19.9	22.5	39.9	17.6	152
total doctor	22.8	21.8	38.0	17.3	345
Lawyer					
oral	20.8	17.9	36.9	24.4	80
written	5.6	33.3	33.3	27.8	3
total lawyer	20.1	18.6	36.7	24.6	83
Other					
oral	23.3	21.7	36.1	18.9	31
written	17.1	18.6	46.1	18.1	24
total other	20.1	20.1	41.3	18.5	55
Grand Total	22.0	21.1	38.3	18.6	483

tions in which each measure accounted for 100% of the question is given in Table 10.

TABLE 7. Comparison of average question components and average question words

	Question	Matching	Translatable	Stop	Untranslatable
Average Component	5.28	1.16	1.11	2.02	.98
Average Number words	5.81	1.35	1.46	2.02	.98
Average Component Percentage		22	21	38	19
Average Word Percentage		23	25	35	17

TABLE 8. Average percentage of Q covered by measures by format

Format	*E*	*C*	*Tr*	*Untr*	*N*
Oral	24.0	44.4	80.8	19.2	304
Written	19.2	41.1	82.0	17.9	179
Total	22.0	43.0	81.3	18.6	483

TABLE 9. Average percentage of Q covered by measures by profession

Profession	*E*	*C*	*Tr*	*Untr*	*N*
Doctor	22.8	44.6	82.6	17.3	345
Lawyer	20.1	38.7	75.4	24.6	83
Other	20.1	40.2	81.5	18.5	55
Total	22.0	43.0	81.3	18.6	483

TABLE 10. Number and percentage of 100% measures

	N	*E*		*C*		*Tr*		*Untr*	
Profession		No.	%	No.	%	No.	%	No.	%
Doctor									
oral	193	17	9	37	19	96	50	10	5
written	152	10	7	33	22	74	49	–	–
Lawyer									
oral	80	8	10	14	18	31	39	2	2
written	3	–	–	–	–	1	33	–	–
Other									
oral	31	–	–	2	6	11	35	–	–
written	24	1	4	1	4	8	37	–	–
Total	483	36	7	87	18	221	46	12	2

• Discussion

The amount of mediation required between user question and system response is the opposite of system responsiveness. That is to say, mediation is required to the extent that the system is not responsive. The mediation can be conducted by the user himself; when he recognizes that the system is not responsive, he modifies the question until the system responds. Traditionally mediation has been conducted by a trained intermediary; he has some knowledge of the user, negotiates the question, determines the limitations on the search, and translates the question terminology into system terminology in accordance with his knowledge of system components.

This study failed to demonstrate differences in system responsiveness according to user profession. The content of doctors' questions may have been different from the content of lawyers' questions but there was little difference in the correspondence of question components to system components. This suggests that, for MDS at least, there is no need to differentiate between medical "specialist" (doctor) requests and medical "nonspecialist" (lawyer) requests to improve responsiveness.

The failure to demonstrate differences in system responsiveness according to the format (oral, written) of the request casts uncertainty on the need for question negotiation. That is, there is doubt if the purpose of negotiation is to increase the proportion of the matching component. Table 5 shows that although oral (negotiated) questions were higher than written (nonnegotiated) in M, this difference was not due to either T or U, but to S. This indicates only that written questions tend to have more nonsubstantive words, as might have been expected. If, on the other hand, the purpose of negotiation is to increase the precision of system responsiveness, this purpose can perhaps best be achieved by the user reformulating his question as a result of system response.

SEARCH LIMITATIONS

The finding that at MDS there were seldom limitations imposed on searches might be interpreted to cast doubt on the need for including such limitations in systems for direct man-system interaction. However, there are other considerations: In the case of written questions (i.e., questions that were not negotiated) the user could not know that limitations might be useful or necessary because he ordinarily did not have a search request form to guide him. In the case of oral questions (i.e., questions that were negotiated) limitations might not have been asked for by the searcher or they might not have been recorded. Limitations would tend not to be asked for if (1) the question were estimated by the searcher to require only a short search for an answer (author search, single reference); (2) the usual needs of a particular class of user (profession or some breakdown of profession) were known to the searcher; and (3) the usual needs of a particular user were known. The experience and bias of the searcher therefore influences the recording of search limitations.

Cases in which limitations on search were stipulated were so few that it was not thought useful to differentiate them into professions or formats. The sparse data available showed that there was a tendency to stipulate English language articles. When the time for completion was specified, the requestor wanted the product in less than 1 week.

In comparison, Kronick (5) found that "English only" was specified in all but 19 of 700 instances at the Cleveland Medical Library. The difference was probably because a higher percentage of the MDS requests were for "in depth" searches; i.e., everything available was wanted. Kronick also found that 57% of his requests were for recent (up to 2 years) material, 29% for 2-5 years, 5% for 5-10 years, and 3% for 10-20 years. Analysis of time period to be covered in the present study suggests concurrence with his figures in those instances when time period was specified.

Thus limitations on searches seem to be more often implicit than explicit. In other words, stipulations are implied by the kind of question and who posed it, rather than being detailed. In considering direct man-machine interaction, the importance of designating search limitations for cutting the size of the file to be searched is considerable. The findings of the present study suggest that the user does not ordinarily consider the importance of delimiting the search; therefore, the system would have to provide a specific checklist or other direct procedure for eliciting such information.

Measures of System Responsiveness

Equality is the criterion of absolute match between user and system terminology, just as would be necessary for a computer to recognize match. In a system with just one component, such as a subject heading list, mediation is required for all question components except the matching one.

The results show that between a fifth and a quarter of the question terms for this study can be matched by MeSH without modification of MeSH or special training of the user (Tables 8 and 9). In addition, in 7% of the questions, the entire question could be matched (Table 10). Because these were "real" questions, these measures may indicate what percentage of response a user would obtain from a machine system using MeSH; however, it is unknown whether he would formulate his question to a machine system in the same way as he now formulates a reference question to a human searcher (especially if the machine vocabulary were provided to the user).

Users' terminology can be considered an indexing language in the sense that question terms are formulated to indicate a body of information. It is therefore of interest to compare the correspondence between user and system terminologies in the present study to studies of the correspondence between and among indexing languages.

Users' terminology is about as equal to MeSH in the present study as MeSH is to LC in a study by Brooks and Kilgour (6). They found 37% of subject headings exactly the same in MeSH and LC. Adjusting the equality figures in the present study for exclusion of stop-list words (which do not appear in subject headings) gives an average of 36%. Therefore users' terminology is as close to the indexing language as these two indexing languages are to each other. This raises the question of the generalizability of the findings in the present study. The equality between indexing languages is also considered important because of studies of compatibility among indexing languages (Schultz (7), Painter (8), Hammond and Rosenborg (9). The word "compatibility" was chosen in the present study to be consistent with usage in these compatibility studies.

Compatibility is a measure of what percentage of users' ideas are dealt with in the system. It includes both matching and translatable question components and reflects the increase in system responsiveness obtained with a two-component system. Compatibility averaged 43% (Tables 8 and 9); 18% of the questions were 100% compatible (Table 10).

When the compatibility figures are adjusted for exclusion of the stop-list, a figure of 70% is found for comparison with other studies. Schultz (7) found that three-quarters of the subject heading list for indexing the meeting papers of the Federation of American Societies for Experimental Biology (FASEB) was accommodated by both MeSH and NIH Research Grants Index authority list. Similarly, Brooks and Kilgour (6) found 79% of MeSH subject headings "adequately covered" by LC. These comparisons suggest that users' vocabularies are as similar to indexing languages as indexing languages covering the same subject areas are to each other.

In evaluating any two languages, a comparison of equality and compatibility gives a criterion for the need to expand cross-references to gain greater equality. Analysis for type of translation needed would show what sorts of programming are required for direct user-machine interaction. If the findings of the present study are borne out by future analyses of the similarity between MeSH and user terminology, a need would be established for developing machine programs that deal with simple inversions and word variants; also a need for expanding dictionaries of cross-references to deal with synonyms. The results also suggest that compounding (one to more than one transformation) and generic-specific relationships, although difficult to handle, are statistically less frequent problems than other types of translations.

Translatability (as used in this study) indicates the further refinement which might be brought about by adding a stop list to the system ($M+T+S$). The average question was 81% translatable (Tables 8 and 9); almost 50% of the questions were found to be 100%

translatable (Table 10). This means that, given a system with a vocabulary list, appropriate algorithms for translation, and a stop-list, half the time it would respond fully to a user's question; or, looking at it the other way, 80% of the average questions could be handled without mediation.

Addition of the *M* and *S* percentages shows that, on the average, 60% of questions could be handled by a system having just a vocabulary list (such as MeSH) and a stop-list.

The remaining question component is the untranslatable fraction; 19% of the average question remained untranslatable (Tables 8 and 9); only 2% of the questions were 100% untranslatable (Table 10).

Adjusting the untranslatability figures to exclude the stop-list terms gives an average figure of 30% for comparison with other studies. Schultz (*7*) found that 10% of the FASEB terms was not accomodated by at least one of the other vocabularies (MeSH, ASTIA descriptors, and the NIH Biomedical Sciences Dictionary). Brooks and Kilgour (*6*) found 15.8% of LC subject headings used at the Yale Medical Library "unmatched" in MeSH. Hammond and Rosenborg (*9*) found 10.9% of the Atomic Energy Commission (AEC) terms had no equivalents in the ASTIA descriptor list. It is difficult to determine whether the differences between the present study and the other three are due to different translation rules or to absolute differences between user vocabularies and MeSH on the one hand, and between systems' vocabularies on the other. A major proportion of the difference may be attributed to the larger vocabulary of natural language as compared with controlled vocabularies. In any case, the untranslatable fraction would appear a major area for system improvement.

Whether the quantitative values of the various measures are adequate for any one system must be left to the judgment of each system designer. Empirical studies of the tolerance of users to different quantitative values are required. However, the method of measuring developed here may be of temporary help for evaluating indices or thesauri according to user vocabularies and for making changes in response to user terminology.

Terminology Analysis

Since the reliability of the present study depends to a great extent on the criteria used for analyzing terminology, some comments are in order on the method used in deriving question components. Similarly comments are required about factors which may have influenced the validity of the findings.

The average question components are shown in Tables 3, 4, and 5. These average findings should be interpreted as conservative figures because of the rules used in analysis. (See Methods: Terminology Analysis). Thus, a term was only considered a match if it was identical to the MeSH entry. For example, a hyphen in the question term which otherwise matched MeSH was enough to put the term in the translatable category (e.g., oculomotor, question 3). Because of the left to right analysis of the questions, a query about "myelogenous leukemia" was not considered a match for "leukemia" but a translatable for "leukemia, myelogenous."

Another reason that the average component figures should be interpreted as conservative is that, although MeSH was chosen as the standard for medical vocabulary, the users were not addressing MEDLARS. They were addressing the holdings of the Library of the College of Physicians and specifically MDS. There were questions that were essentially architectural (question 27). There were questions that required textbook material for the answer, such as pictures of normal anatomy (questions 191–194). Presumably if the questions not suited to MEDLARS had been eliminated, the proportion of untranslatable words would have been lower.

Again, if the 16 Spanish language questions had been eliminated, the proportion of untranslatable words would have been lower. It might be expected that foreign language questions would inflate *T* in comparison with *M*. This was not necessarily true since much of medical terminology is international. In fact, some of the requests from Spanish speaking countries that were in English were less likely to match than the ones in Spanish; for example, the appearance of "different diagnosis" in question 82, and "nourse children" in question 81. However, the system vocabulary would not ordinarily be expected to have foreign language stop words as "de" and "la" which appear on the list of most frequent stop words (Table 2); the stop-list was distinctly affected by the foreign language requests.

The approach in deriving the translatable list was also conservative: when there was doubt, the word was considered untranslatable. Of the categories of translatable terms (Appendix E), synonyms, compounding and generic-specific relationships proved most difficult to handle. Many synonyms were abbreviations; others were provided by Anglo-Saxon forms where the Latin or Greek form appeared in MeSH or the other way around. The usual cross-references in a dictionary or authority list would not be expected to have all of these. Instances of compounding occurred mainly with terms such as "tympanoplasty" which was translated to "tympanic membrane" and "plastic surgery."

Generic-specific relationships were most difficult; the difficulty lay in determining a rule for a cutoff in the hierarchy. For example, "ablation" is specific to "surgery," "arteriospasm" specific to "vascular diseases." For this study, the hierarchical distance in these examples was considered too great for useful inclusion as translations. Therefore, both "ablation" and "arteriospasm" appear on the untranslatable list (Appendix H).

The problem of what level of generality or specificity should be the cutoff point in translation is illustrated by the Hammond and Rosenborg study (*9*) of the convertibility between the ASTIA descriptor list and AEC dictionary. They assigned six different categories to what has been termed translation here. Of these, four dealt with the generic-specific problem. However, the findings of the present study suggest that the generic-specific relation represents only a small proportion of the total amount of translation required. Perhaps a solution to this problem is not as urgent as others.

Singular or plural forms comprised a large proportion of the word variant type of translation and invite comparison with a recent study by Bloomfield (*10*). He studied indexing of singular and plurals in Webster's Unabridged Dictionary, *Chemical Abstracts, Nuclear Science Abstracts,* and an IBM KWIC index and found marked inconsistencies. Since conversion to singular or plural form represented a large category of translations in the present study, medical terminology may be added to Bloomfield's list. Although MeSH was not studied formally for singulars and plurals in the present study, MeSH was found to be inconsistent. Sometimes one is used and sometimes the other, although no instances of both forms appearing were noted. Some of the reasons for inconsistency in use of singulars and plurals became clear as the analysis proceeded. One reason is apparently to differentiate homographs; for example, "joint" may be an adjective or a noun but "joints" is clearly a noun and, in a medical context may be taken to refer to arthrology. Another reason is the persistence of Latin and Greek forms in medical terminology. The plural of carcinoma may be either carcinomata or carcinomas. There seems to be a tendency for the MeSH entry to have been constructed to avoid this problem, although no tabulation was made. Question plurals revealed both English and Latin or Greek forms (including a few incorrect ones). Another source of the singular being used in a question occurred when just a single instance was specified. A final class was those terms appearing in a question as singular because they were parts of nominal compounds (e.g., question 208, mesenteric artery thrombosis).

The stop-list was also derived conservatively in the sense that a doubtful word was considered untranslatable unless it was neither substantive nor appeared more than two times. Therefore, the untranslatable list includes many words which might be considered candidate stop-list words if they were to appear with significant frequency. The untranslatable list also includes substantive terms which did not appear in MeSH because they were too specific ("arteriospasm") or because they were not appropriate to MeSH coverage (*Lupinus*).

• Conclusions

Doctors' and lawyers' question terminologies were not shown to differ significantly in correspondence to medical system terminology; this suggests there may be no need, when planning for direct user-retrieval system interaction, to differentiate between "specialist" and "nonspecialist" terminologies when different client groups are trained to approximately the same level in their respective disciplines. Oral and written questions were not shown to be significantly different in question components and hence system responsiveness; this casts doubt on the traditional need for question terminology negotiation. Specifications about searches were shown to be infrequent, which suggests the need for explicit methods of delimiting searches in user-retrieval system interaction in order to limit the size of the file to be searched.

About 50% of the questions in this study were 100% translatable ($M+S+T$); they could be transacted with no intermediary given the subject heading list, stop-list, and appropriate translation rules. About 80% of the average question was translatable. The untranslatable 20% would require either a human intermediary or a complex machine program for terminology negotiation to fill the communication gap between user and retrieval system.

• Acknowledgments

The author is grateful to Elliott Morse, Librarian of the College of Physicians of Philadelphia, and Walter Bethel, Director of the Medical Documentation Service, for making the reference questions available for study. Special thanks are due to Mrs. Claire K. Schultz, Institute for the Advancement of Medical Communication, for continuing advice and criticism. Dr. Richard Orr, Institute for the Advancement of Medical Communication, and Roger Sisson, Wharton School of the University of Pennsylvania, gave valuable suggestions. James McGuinnis, Computer Department of the Drexel Institute of Technology, advised on the appropriate statistical techniques.

References

1. Kessler, M. M., E. L. Ivie, and W. D. Mathews, The M. I. T. Technical Information Project—A Prototype System, *Proceedings of the American Documentation Institute,* 1:263–268 (1964).
2. Council on Library Resources, Inc., *Automation and the Library of Congress,* Library of Congress, Washington, D.C., 1963.
3. Overhage, C. F. J., and R. J. Harmon (eds.), *Report of a Planning Conference on Information Transfer Experiments,* Sept. 3, 1965, The M. I. T. Press, Cambridge, Mass., 1965.
4. Berul, L., *Information Storage and Retrieval: A State-of-the-Art Report,* AD 630–089, Auerbach Corp., Philadelphia, Pa., 1964.
5. Kronick, D. A., Varieties of Information Requests in a Medical Library, *Medical Library Association Bulletin,* 52 (No. 4): 652–669 (1964).
6. Brooks, B., and F. G. Kilgour, A Comparison of Li-

brary of Congress Subject Headings and Medical Subject Headings, *Medical Library Association Bulletin,* 52 (No. 2): 414–419 (1964).

7. Schultz, C. K., Guide to Current Terminology in Biomedical Research, *Federation Proceedings,* 24 (No. 4): 960–963 (1965).
8. Painter, A. F., *An Analysis of Duplication and Consistency of Subject Indexing Involved in Report Handling at the Office of Technical Services,* PB 191501, U.S. Department of Commerce, Washington, D.C., 1963.
9. Hammond, W., and S. Rosenborg, *Experimental Study of Convertibility between Large Technical Indexing Vocabularies,* Tech. Rep. IR–1, Datatrol Corp., Silver Spring, Md., 1962.
10. Bloomfield, M., A Study of Singular and Plural Words as Index Terms, *Proceedings of the American Documentation Institute,* 3:201–205 (1966).

The Relationship of Natural and Social Sciences to Social Problems and the Contribution of the Information Scientist to Their Solutions*

Social problems have multiple causes, and their solutions accordingly require a multidiscipline approach, which is facilitated by the fact that technology, the natural sciences, and the social sciences are closely interrelated (a point of view that is making itself increasingly felt in educational theory). The deterioration of the inner city is an example of a typical complex social problem that will yield only before such a unified attack. Solutions to social problems have been suggested by findings from such varied fields as astrophysics, sensory psychophysics, and population studies, as well as the more theoretical social sciences, whose influence can be seen in their application to problems of urban development. Although information specialists have hardly yet developed a full-fledged body of knowledge, they can contribute much towards the solving of social problems by: (1) organizing and disseminating information to those broad-gauged individuals and groups that are working on problems that defy solution by a fragmented approach; (2) controlling the rapidly mushrooming body of pertinent technical literature; (3) developing periodical indexes for the field of social welfare; and (4) assisting in the compilation of state-of-the-art papers on social problems, basing their work on a value-oriented extracting technique derived from a model created by the National Association of Social Workers.

JOE R. HOFFER

National Conference of Social Welfare
Columbus, Ohio

• Introduction

There has been a growing recognition by theoretical and applied social scientists that interdependence is a condition of modern life. Accordingly, the major assumptions of this paper are:

1. This interdependence extends beyond the social sciences and includes the physical and the biological sciences.
2. Social problems result from a variety of causes, and to obtain real understanding, knowledge needs to be drawn from many sources.
3. This knowledge is not restricted to the purview of a single discipline or field, nor can it be divided into mutually exclusive categories.

The present paper will explore the premise that information science and information specialists can make a major contribution to the solution of basic social problems by collecting and integrating pertinent knowledge from the physical, biological, and social sciences, and by relating it directly to selected critical areas. These assumptions will be discussed under the following major headings:

1. What are some of the basic social problems?
2. What is the relationship of the physical and biological sciences to the social sciences, and the relation of the three to social problems?
3. What are some possible contributions of the information scientist?

• What Are the Basic Social Problems?

For our purposes, the basic social problems are those that are fundamental and universal—problems whch exist in many communities and countries. Just as different scientific and professional disciplines emphasize different aspects of the whole person to arrive at a "theory of man" (philosophical, theological, biological, psychological,

[*] Presented at the Central Ohio Chapter, American Documentation Institute, Columbus, Ohio, April 21, 1966.

psychoanalytical, sociological, etc.), it is understandable that these professions and scientific disciplines will emphasize different social problems.

In the words of Arthur Blum, "Various authors have described us as a lonely crowd (*1*), growing up absurd (*2*), in an insane society (*3*), composed of status seekers (*4*), exurbanites (*5*), organization men (*6*), and the invisible poor (*7*). These writings describe systems of extensive distortions in our present social functionings. What is of additional interest about this collection of books is the variety of backgrounds and disciplines represented by the authors, indicating increasing concern with common problems within a number of different fields" (*8*). The monumental rise in psychoses, neuroses, alcoholism, crime, delinquency, poverty, divorce, and suicide, to name only a few of our social problems, increasingly testifies to the extent and variety of the social pathology which confronts us today.

• What Is the Relationship of the Physical and Biological Sciences to the Social Sciences, and the Relationship of the Three to Social Problems?

The solution of the social problems of our society requires the productive interaction of many disciplines —the natural sciences, the social sciences, engineering, and management.

The findings of the natural sciences have a profound effect upon social science, social welfare, and social problems. This is as it should be, for man himself was first scientifically investigated as a physical entity. Nevertheless he is a social being, and as such is the basic unit of concern in the study of collective behavior.

Interrelationships are particularly easy to see in the influence of chemistry, pharmacology, and medicine upon areas of concern to public health. Take venereal disease, for example, with its host of accompanying social complications. It can now be effectively brought under control thanks largely to a medical breakthrough, namely the discovery of antibiotics. An even vaster problem, mental and emotional illness, which could be said to affect directly or indirectly almost every area of social malfunctioning, now shows some sign of yielding to certain chemical compounds known as tranquilizers and antidepressants. The application of DDT and similar insecticides has made possible the effective eradication of pest-borne diseases (such as malaria) which formerly were endemic to many areas of the world.[1]

In another area, technology (the child of the natural sciences) has made possible, in countries such as ours, so great an abundance of the necessities of life and so great an increase in national resources that we now, in all seriousness, propose to abolish poverty, one of the most ancient and universal of our social problems.

Nor can the body of knowledge of the social sciences escape the impact of technology. Such an increase of wealth affects in many ways (both obvious and subtle) the very structure of society, thus providing new raw materials for the economist (who studies the implications of the distribution of this wealth), for the sociologist (who studies its effects on social class and stratification), and for the political scientist (who studies the significance of this social change for governmental institutions and processes). The findings of such inquiry is of profound significance to the theory of human collective behavior.

If scientific breakthroughs can solve social problems, they can also create them. Progress in public health, by lowering the death rate, acts as a primary factor in the population explosion; and we are all too familiar with the social and political consequences of certain well-known discoveries in nuclear physics. The computer, with its impressive potential for the speedy and accurate storage, integration, control, dissemination, and implementation of data may also turn out to be the villain in its forthcoming role as the boss of the automated factory, which some economists and labor leaders see as the root of large-scale future unemployment. Social welfare, and especially the field of recreation, view automation as the source of a possible superabundance of leisure, which, for healthful living, they must help to fill with meaning.

Interdisciplinary Relationships and Education

Regarding the relation of science to the social sphere, Derek J. deSolla Price (*9*) believes that both government and scientists are interested in ensuring that science be promoted for the good of society and the nation: "A scientist," insists Price, "does not need political motivation to be conscious of the social relations of science."

This interdependence is reflected in the field of education, whose leaders are aware that, as C. P. Snow has said, "the purely scientific education is incomplete, but a purely non-scientific education is also incomplete. (*10*)."

The schools of the future will incorporate this concept into the curriculum. A recent article in the New York Times reports on a study recommending "that all students be offered a new set of general education courses in their senior year in which they would relate the liberal arts to such specific areas as urban renewal, the development of new states, the problems of the public bureaucracy, and the philosophy of science . . . that every student, in addition, should be required to take a one-term course in economics, sociology, government, anthropology, or geography . . . that all students should take a two-year mathematics-physics, or mathematics-biology sequence (*11*)."

It is also evident that our educational institutions must give greater attention to the developing of broad-

[1] However, the new insecticides are, in turn, producing problems, for the Department of Health, Education, and Welfare and the Department of Agriculture are both becoming quite concerned about the residual buildup of insecticides both in the soil and in ground water.

gauged leaders. In the words of John W. Gardner, Secretary of the Department of Health, Education, and Welfare, "Leadership is dispersed among a great many groups in our society. . . . Nothing should be allowed to impair the effectiveness and independence of our specialized leadership groups. But such fragmented leadership does create certain problems. One of them is that it isn't anybody's business to think about the big questions that cut across specialties—the largest questions facing our society (*12*)."

What is a Social Problem? A Typical Example

As one studies the current activity on a specific front—for example, urban problems—one is led, as Rouse has expressed it, "to a depressing conclusion . . . there is absolutely no dialogue in the U. S. today between the people who have developed knowledge about people—the teachers, ministers, psychiatrists, sociologists—and the people who are designing and building our cities (*13*)." He believes that we are not asking the right questions, and so we are not getting the right answers.

The deterioration of the inner city is an example of a social problem condition that must be considered in urban development. The inner city refers to "a zone of land that circles the central business district of the metropolis, extending outward toward the city boundaries (*14*)."

Northwood (*15*) defines four major concepts in sufficient detail to permit forward movement: social conditions, social problems, social work, and the inner city. In his analysis, he catalogs an impressive list of conditions associated with the core of the city under three major headings: (1) land use in the inner city; (2) people of the inner city; and (3) social control of the inner city.

Northwood's major assignment was to examine these conditions and to identify to what extent, traditionally and currently, organized social welfare has recognized them as social problems appropriate to its work. A secondary assignment was to assess the contribution of social welfare to the solution of these social problems.

There is ample evidence that social welfare and the social work profession have attempted to ameliorate such social conditions and problems in a wide variety of specific programs and services over a long period of time. There is also ample evidence that these conditions and problems are persistent and hardy, and will require new and revolutionary approaches and services.

The evidence would favor, in Cohen's words, "broad programs of social reconstruction as against specific programs for specific social problems. This economic and physical planning cannot be separated from social and psychological planning (*16*)."

The Concern of the Physical Sciences

There are some significant examples of the application to social problems of methods developed by the physical sciences. The Pennsylvania Department of Public Welfare is working with Dr. Leon Stegg, astrophysicist and manager of the Space and Missile Technology Center of the General Electric Company, with a view toward using the systems analysis approach to the solution of human problems, particularly in public welfare. Similar work has been done in California, and a bill has been introduced in Congress to give financial support to the states for further experimentation in this area.

Methods of sensory psychophysics have been used to gauge the intensity of opinions and attitudes (*17*). Experiments in a dozen laboratories have shown how procedures developed for the scaling of sensory attributes such as brightness and loudness can measure human reactions to many forms of nonmetric stimuli. Stevens tells how this procedure (called "magnitude extimation") has been used to assess the consensus concerning intensity or degree for such variables as strength of expressed attitudes, pleasantness of musical selections, seriousness of crimes, and other subjective dimensions for which the stimuli can be arrayed only on nonmetric or normal scales.

The Concern of the Biological Sciences for Social Problems

Guhl (*18*) in his article "Sociobiology and Man," reviews current knowledge of the biological basis of human sociality, and suggests that sociobiologists need to be aware of current information in the human social sciences.

To be specific, let us turn to a realm of social problems in which changes in value orientation on the part of the populace have been occurring, namely family planning. As Dael Wolfle suggests, "The fundamental problem is people. Whatever we do to increase food supplies, conserve water, improve land management, or curb pollution merely postpones for a few years the day of catastrophe unless we stop increasing the number of hungry mouths. . . . (*19*)." Within the past few years remarkable progress has been made in this field, despite the intense values of what might have been a majority of the population in the beginning. Writing in the *New York Times,* Ambassador Bowles states:

> India has embarked upon the world's most ambitious control program. A program which one Indian official describes is at the very center of planned development. The number of family planning clinics, which totaled 144 seven years ago, has increased to 8,504—most of them in rural areas. By 1966, some 4,000 new clinics will have been added. 20 million posters and 60 million pamphlets have been distributed on the subject of family planning. Moreover, there is some evidence, admittedly uncertain, that the effort is already showing results. National birthrates, which averaged 48 per thousand in 1951, have dropped to 40. Moreover, in Bombay, where a vigorous program has been in effect for several years, the rate is down to 28 per thousand. In rural areas, intensive pilot programs have resulted in a decline in the birthrate by as much as 30%. Indian planners and U. S. advisers, working through

> the Ford Foundation, feel there is a reasonable chance that India's population may begin to balance out in the 1970's, with a birthrate of 25 per thousand and a death rate of 20 (*20*)."

A dramatic illustration of the implications of the population growth in urban areas is an estimate, made by the United States Municipal News and based on U. S. conditions, which indicates that "Every 1000 new people in the metropolitan area require: 4.8 elementary school rooms; 3.6 high school rooms; 8.8 acres of land for schools, parks, and play areas; an additional 100,000 gallons of water per day; 1.8 new policemen; 1.5 new firemen; 1 additional hospital bed; 1000 new library books; a fraction of a jail cell; sewage and treatment for 170 pounds of organic water pollutants per day (*21*)."

The Social Sciences: Their Relations with the Natural Sciences and Their Concern for Social Problems

Talcott Parsons (*22*) provides some basis for analyzing interrelations among the various disciplines by examining the formal preparation of social scientists in fields other than their own. He suggests that there is a historical and logical basis for dividing intellectual disciplines into the natural sciences, the social sciences, and the humanities. By these gross measures, the social sciences are largely independent of the natural sciences and mathematics,[2] but have closer ties to the arts, the professions, and the humanities.

An excellent example of the relationship of the social sciences to a key social problem area is found in the field of urban development, to which we referred earlier in this section. Lawrence K. Northwood believes that "most information about the city and its subareas is found in studies of the social scientists. Such studies vary widely among the academic disciplines. The economic geographers originally stressed the physical habitat of man, but have moved over to cite human ecology; the anthropologists visited agrarian communities overseas, but now probe into industrialization and urbanization, no matter where their locus may be; political scientists have tended to examine the formal and informal boundaries of power and its social structures, investigating international, state, and local questions; rural sociologists and adult educators traditionally have been concerned with small-town social action systems (*25*)."

As for another example, social welfare and the social work profession draw heavily upon other fields and professions. To quote Taber and Shapiro "evidence of borrowing knowledge from other fields was found in the use of recognized authorities, concepts, and theories for exposition or interpretation (*26*)." To quote Ruth Butler, "It is evident that collaborative work with related sciences and professions will be needed to secure the content required for understanding of each knowledge area recommended as a desirable objective (*27*)."

• What Contribution Can the Information Scientist Make to the Solution of Social Problems?

Although there is little evidence that information science has achieved the full status of a profession, the growing literature on information and its documentation suggests that this goal is not too far off. It appears that a recognizable common body of special knowledge will evolve from such present separate and independent activities as library science, documentation, information storage and retrieval, linguistics, machine translation, and information systems engineering. Furthermore, the literature and the growing complexities of dealing adequately with knowledge leave little doubt that a second requirement of a profession will be met, namely, a recognized task for its members to perform.

Mooers (*28*), Cuadra (*29*), Heilprin (*30*), Slamecka (*31*), Kent (*32*), and Crosland (*33*) have analyzed the development of this new profession and have attempted tentative definitions. Perhaps the most useful one for our purposes was reported by Heilprin, namely "the science that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. The processes include the origination, dissemination, collection, organization, storage, retrieval, interpretation, and use of information. The field is derived from or related to mathematics, logic, linguistics, psychology, technology, operations research, the graphic arts, communications, library science, management, and some other fields (*34*)."•

Since the focus of this paper is on the possible contribution of the information scientist to the solution of social problems, it is necessary to provide an operational definition for the practitioner in this profession. The information scientist has been defined as follows: "One who studies and develops the science of information storage and retrieval, and who devises new approaches to the information problem, who is interested in information in and of itself (*34*)."

It is my conviction that an information scientist with an orientation in the physical, biological, and social sciences can make a vital contribution to researchers,

[2] As for the increasing role of mathematics in social sciences, Abraham Kaplan writes that "A troubling question for those of us committed to the widest application of intelligence in the study and solution of the problems of men is whether a general understanding of the social sciences will be possible much longer. Many significant areas of these disciplines have already been removed by the advances of the past two decades beyond the reach of anyone who does not know mathematics; and the man of letters is increasingly finding, to his dismay, that the study of mankind proper is passing from his hands to those of technicians and specialists . . ." (*23*).

This suggests that to attain mathematical competence the social sciences should consider a marriage with mathematics rather than some other less systematic arrangement. The *American Sociologist*, in its November 1965 issue, reports that "The application of mathematics and logic in sociology appears to have reached the point that active research and teaching require some background in mathematics, and specialized positions in sociology departments are needed which only can be filled by persons with mathematical training" (*24*).

administrators, and policy groups in solving our pressing social problems.

The Control of Technical Literature

An interesting presentation of the acceleration of our time has been prepared by J. Lewis Powell (*35*). He compresses the 50,000 years of mankind's recorded history into 50 years, and on this basis develops the following chronology:

1. Ten years ago, man left his cave for some other kind of dwelling.
2. Five years ago, some genius invented the first writing.
3. Two years ago, Christianity appeared.
4. Fifteen months ago, Gutenberg developed the printing press.
5. Ten days ago, electricity was discovered.
6. Yesterday morning, the airplane was invented.
7. Last night, radio.
8. This morning, television.
9. The jet airplane was invented less than a minute ago.

We might add, earth-orbiting occurred 30 seconds ago, a moon shot may occur within the hour, and interplanetary vacations may occur tomorrow.

Consider this time-table in relation to the following definition of science, which identifies it totally with its documentary output: "Science is that which is published in scientific journals, papers, reports, and books. In short, it is that which is embodied in the literature (*9*). These facts, together with the simple observation that recorded knowledge accumulates through the years, whereas the rate at which it can be read by any person remains constant, have profound implications, not only to scientists and administrators, but to information specialists, namely: that the technical literature of all sciences and professions is accumulating at a rate which has been compared to that of geometric progression, and that the heavy burden of introducing order into this potential chaos lies squarely upon the shoulders of librarians, documentalists, and information specialists.

Generalization versus Specialization: the Role of the Information Scientist

The increasing volume of publications means inevitably that the expert tends to confine his attention to an ever-narrowing area, while the literature tends to become more and more scattered. K. William Kapp, for example, in his book entitled *Toward a Science of Man in Society*, highlights the problems of generalization and specialization: "Systematic scientific inquiry in the social sciences today is marked by a curious contradiction. On one hand, we are witnessing a rising demand for intellectual cooperation and integration, which finds expression in various interdisciplinary endeavors and cooperative ventures by scholars from different disciplines; on the other hand, the traditional compartmentalization of the social sciences has continued, and is vigorously defended on the ground that specialization is the prerequisite for all creative work in scientific inquiry, as indeed in all fields of human endeavor (*36*)."

The suggestion that we combine the various sciences into a single body of knowledge focused on a selected social problem raises some real difficulties for the information specialist, namely that it appears evident that there is a need for a form of common integration, semiautomatic in nature, that will cross the various disciplines and feed its findings back to the individual researcher. There are two principal obstacles that must be overcome before such a common integration can be produced. One is the increasing volume of publication and the other is the present incompatibility of the physical, biological, and social sciences.[8]

Information science can make a definite and related contribution by organizing and disseminating information to those broad-gauged and generalist individuals and groups that are working on social problems that defy solution by a fragmented technique or approach.

Some Specific Contributions Towards the Solution of Social Problems

The general function of the information specialist is the organization of the literature, a function which can act as a force toward integration and synthesis. However, one concrete and major contribution information specialists and librarians could make toward the resolving of social problems would be to strengthen the basic bibliographic tools of the professions which must cope with them. A student, researcher, practitioner, or administrator who wishes to locate articles in which he is interested might have to consult as many as 16 different indexes and indexed abstract journals, none of which really specializes in social welfare, or claims to list extensively papers dealing with it.

Another specific contribution that information specialists can make is to assist researchers in compiling state-of-the-art papers on selected social problems. To increase the usefulness of available knowledge, a model (*38*) developed by the National Association of Social Workers is suggested. In brief, the model is value oriented, and emphasizes the gaps between the ideal objectives and the actual operations.

The model stresses the following major considerations:

1. Definition and etiology of the problem.
2. The societal norms and values, and the assumed scientific and professional norms and values affecting the problem.
3. The current programs actually dealing with the problem, and the consequence of continuing these programs.
4. The ideal or the social change objective.

[8] As for the latter, social science data do not resemble physical and biological data sufficiently to be comparable to them. Foskett suggests that social scientists have approached modern retrieval systems with caution because librarians have assumed that there is little difference in these various data. He stresses the fact that "what distinguishes the social sciences, perhaps, is the extent to which subjective attitudes and imprecise terminology appear in the literature, and the masterful manner in which some scholars dispose of their opponents" (*37*).

5. The relationship between the actual and the ideal: Identification of the gap between them, sources of resistance to, and support of, the closing of this gap, action priorities for scientific and professional groups, and theory and research needs for attaining necessary knowledge and programs.

As for the application of this technique to the literature of a specific social problem, one would first, of course, have to assemble the pertinent documents. Because of the multifaceted quality typical of most social problems, an information specialist or librarian conducting such a literature search should be able to bring a wide spread of materials, from many sources and professions, to focus on his topic. After a working bibliography has been compiled, he might proceed to apply the model by extracting key sentences and paragraphs according to its five rubrics. Not every article would have material to match each breakdown, but negative results would be as meaningful in their way as positive ones. The researcher—sociologist or social worker—would now be ready to make the final selection and to write the article.

• Conclusion

The achievement of the goal of interdependence is complicated by the fact that each profession and field has maintained that specialization is the prerequisite for all creative work in all human endeavor. However, there is a rising demand for intellectual cooperation and integration, a demand which finds expression in various interdisciplinary enterprises and cooperative ventures by scholars.

The last 10 years have brought a new expertise to the administration and nourishing of science and technology for social goals. Information specialists can relate this new body of knowledge to the specialized information and experience which have accumulated over the years. The net result will provide a rich cross-fertilization among the various fields of scholarship and between each of them and the relevant areas of expert organizational knowledge.

References

1. Riesman, D., *The Lonely Crowd,* Yale University Press, New Haven, 1950.
2. Goodman, P., *Growing Up Absurd,* Random House, New York, 1960.
3. Fromm, E., *The Sane Society,* Rinehart, New York, 1955.
4. Packard, V., *The Status Seekers,* McKay, New York, 1959.
5. Spectorsky, A. C., *The Exurbanites,* Lippincott, Philadelphia, 1955.
6. White, W. H., Jr., *The Organization Man,* Simon & Shuster, New York, 1956.
7. Harrington, M., *The Other America; Poverty in the United States,* Macmillan, New York, 1962.
8. Blum, A., Needed: An Interdisciplinary Approach to Social Problems, *Western Reserve University Outlook,* 3(11):6–9 (1966).
9. Price, D. J. de Solla, The Science of Science, *Bulletin of the Atomic Scientists,* 21:2–8 (1965).
10. Snow, C. P., Government, Science and Public Policy, *Science,* 151(3711):650–653 (1966).
11. Hechinger, F. M., The Liberal Arts Find a Defender, *New York Times* (Feb. 28, 1966).
12. Gardner, J. W., The Need for Leaders, *Science,* 151 (3708):(1966).
13. Columbia, Maryland, Being Built by Community Research and Development, Inc., *Architecture Forum* (Sept., 1964).
14. Murphy, R. E., and J. E. Vance, Jr., *in* J. P. Gibbs, Ed., *Urban Research Methods,* Van Nostrand Co., Princeton, N. J., 1961.
15. Northwood, L. K., Deterioration of the Inner City, *in Social Work and Social Problems,* ed. by Nathan E. Cohen (New York, National Association of Social Workers):201–269 (1964).
16. Cohen, N. E., A Social Work Approach, *in* N. E. Cohen, Ed., *Social Problems,* National Association of Social Workers, New York, 1964, p. 385.
17. Stevens, S. S., A Metric for Social Consensus, *Science,* 151:530–541 (1966).
18. Guhl, A. M., Sociobiology and Man, *Bulletin of the Atomic Scientists* (1965).
19. Wolfle, D., Save the World, *Science,* 149:819 (1965).
20. Aptekar, H. H., Value Orientation of Indian Social Work, *Social Work Review,* 11 (1964).
21. Report of the World Health Organization Expert Committee on Metropolitan Planning, *World Health* (Dec., 1964).
22. Parsons, T., Unity and Diversity in the Modern Intellectual Disciplines: The Role of the Social Sciences, *Daedalus* 94:39–65 (1964).
23. Kaplan, A., Mathematics and Social Analysis, in *Game Theory and Related Approaches to Social Behavior,* ed. Martin Shubek (New York, Wiley): 81 (1964).
24. *American Sociologist* (Nov., 1965).
25. Northwood, L. K., *American Sociologist.*
26. Taber, M. and Iris Shapiro, Social Work and Its Knowledge Base: A Content Analysis of Periodical Literature, *Social Work,* 100–106 (1965).
27. Butler, R. M., *An Orientation to Knowledge of Human Growth and Behavior in Social Work Education,* Butterworth, Washington, D.C., 1963, p. 31.
28. Mooers, C. N., The Educational Challenge of Information Science, *Proceedings of the American Documentation Institute* pt. 1, 127 (1963).
29. Cuadra, C. A., Identifying Key Contributions to Information Science, *American Documentation,* 15:289 (1964).
30. Heilprin, L. B., Education for Information Science: Review and Orientation, *in Proceedings of the Symposium on Education for Information Science,* Spartan Books, Washington, D.C., 1965, p. vii, 171. Toward a Definition of Information Science, *Proceedings of the American Documentation Institute,* pt. 2, 239–241 (1963).

31. SLAMECKA, V., On the Nature of Information Science and the Responsibility of Institutions of Higher Education, *in Proceedings of the Symposium on Education for Information Science,* Washington, Spartan Books, Washington, D.C., 1965, p. 91.

32. KENT, A., Information Sciences—Development of an Interdisciplinary Graduate Education Program at the University of Pittsburgh, *Proceedings of the American Documentation Institute,* pt. 1, 111 (1963).

33. CROSLAND, D., Graduate Training in Information Science: Definitions and Developments at the Georgia Institute of Technology, *Proceedings of the American Documentation Institute* pt. 2, 243 (1963).

34. Education for Information Science; Review and Orientation. (Citation of an unnamed conference (Georgia Institute of Technology) by Laurence B. Heilprin) *in Proceedings of the Symposium on Education for Information Science,* Spartan Books, Washington, D.C., 1965, p. viii.

35. J. L. Powell, as quoted by B. Karsh in White Collar Labor, *The Nation,* 93–96 (1959).

36. Kapp, K. W., *Toward a Science of Man in Society,* Nijhoff, The Hague, 1961.

37. FOSKETT, D. J., *Classification and Indexing in the Social Sciences,* Butterworth, Washington, D.C., 1963, p. 18, 36.

38. COHEN, N. E., Ed., *Social Work and Social Problems,* National Association of Social Workers, New York, 1964.

Analysis and Automated Handling of Technical Information at the Nuclear Safety Information Center*

The Nuclear Safety Information Center serves the nuclear community by collecting, storing, evaluating, and disseminating safety information relevant to the design and operation of nuclear facilities. In 1964, after about a year of operation, the information-handling system was computerized in order to increase broadly the scope of the Center's services and enable efficient functioning in the future. Computer programs were developed for the preparation of a bibliography, complete with key-words and personal author indexes, that is issued quarterly and for a program of selective dissemination of information (SDI) that is produced on 5×8 in. cards. These programs and other services of the Center are discussed.

J. R. BUCHANAN and F. C. HUTTON

Nuclear Safety Information Center,
Oak Ridge National Laboratory
and
Computing Technology Center,
Union Carbide Corporation

The USAEC established the Nuclear Safety Information Center (NSIC) at Oak Ridge National Laboratory (ORNL) in March 1963. The Center serves the nuclear community by collecting, storing, evaluating, and disseminating safety information relevant to the design and operation of nuclear facilities (*1*). It was in operation almost immediately after its establishment because the scientists and engineers necessary to the operation of a center, and without which an information center is hardly more than a specialized library, were already on the ORNL staff or available through existing consulting contracts.

The subject of nuclear safety was divided into 19 categories, such as *Accident Analysis,* and technical personnel were assigned on fractional-time basis to study these categories, prepare review articles and reports, answer inquiries, and catalog information. Each reference reviewed by the specialists is indexed according to a system of key words developed by the staff. The key words, title, author, corporate author, and abstract for each document were initially recorded on 5×8 in. cards and duplicate cards were filed under each key word, author, and corporate author.

This system enabled the Center to get into operation immediately and was quite workable as long as information items were not too numerous. After about a year, however, it was decided to computerize the system for use of the IBM 7090s at the Computing Technology Center[1] in order to broadly increase the scope of the Center's services and prepare for future growth without burdening the technical staff with the routine. Two computer outputs were initially planned: the first was a bibliography, complete with key word and personal author indexes, to be issued quarterly; and the second was output in the form of cards for a program of selective dissemination of information (SDI). Both of these outputs are now in operation.

The development of these programs, particularly the SDI, is described and the range of services and organization of the Center are discussed in this review.

• Computer Program Development

A prime consideration in developing the computer programs was to keep the system flexible enough to permit growth of NSIC operation and to make it feasible to extend the system to the work of other information centers without major modifications to the programs.

* Research sponsored by the U. S. Atomic Energy Commission under contract with the Union Carbide Corporation.

[1] The Computing Technology Center, which operates computers and does problem analysis and programming, is operated by the Union Carbide Corporation Nuclear Division for the U.S. Atomic Energy Commission at the Oak Ridge Gaseous Diffusion Plant. F. C. Hutton supervises the Information Retrieval Section of the Information Systems Department of the CTC.

Both these requirements dictated that the programs be fast and that the capacities of the programs be large. Speed was obtained by writing the highly repetitive parts of the programs in symbolic language subroutines usable by the programs, which are written in COBOL. Large capacity was obtained by always being attentive to the amounts of computer core storage required by different techniques. To date, four other centers have used the basic programs.

Four programs make up the system; one will form and update a Master Tape; one will select items from the tape and prepare a bibliography; and one will search the tape in response to questions. The fourth program is used to maintain the key-word file that appears on the front of the Master Tape.

The following information appears on the computer tapes, which are organized in linear fashion so that everything concerning one item on the tape appears together serially; the asterisk indicates that the element is searchable or can be discriminated at this time:

*1. Type, such as reports, journal articles, etc.
*2. Evaluation of contents (as to pertinency)
*3. Category (such as *Accident Analysis*)
*4. Journal abbreviation (ASTM's Codes)
*5. Date
*6. Language
*7. Country
*8. Corporate author
*9. Personal author(s)
*10. Title
11. Description, such as pages, figures, tables
12. Abstract
*13. Key words

Key words are weighted on searching, with an acceptable total weight being specified, and negative weights are permitted. Search elements can be connected on an AND/OR basis. Another form of weighting is used when the references are indexed by assigning an asterisk to the key words of primary importance in each document. These asterisks show in the various printed outputs of the Center.

• SDI Profiling and Operation

In the SDI program, NSIC sends out abstract cards to over 800 members of the nuclear community according to their individual interests. The program was inaugurated in October 1965 on a pilot scale. It gained such an enthusiastic reception from the initial participants that a decision to expand the program was made soon thereafter. Since early in 1966, additions to the program have been made on a routine basis. This is expected to continue for some time.

In setting up the individual profiles, several management level individuals were initially invited to define their specific interests in the field of nuclear safety. Some 50 of the group of 80 responded with the needed information. A profile of interest was then constructed for each participant by assigning the appropriate key words from our indexing vocabulary in a fashion similar to that used in indexing the subject content of a document. Structuring was done in such a fashion as to give a small fraction of false drops and not have the profile drawn so compactly as to eliminate references of true interest.

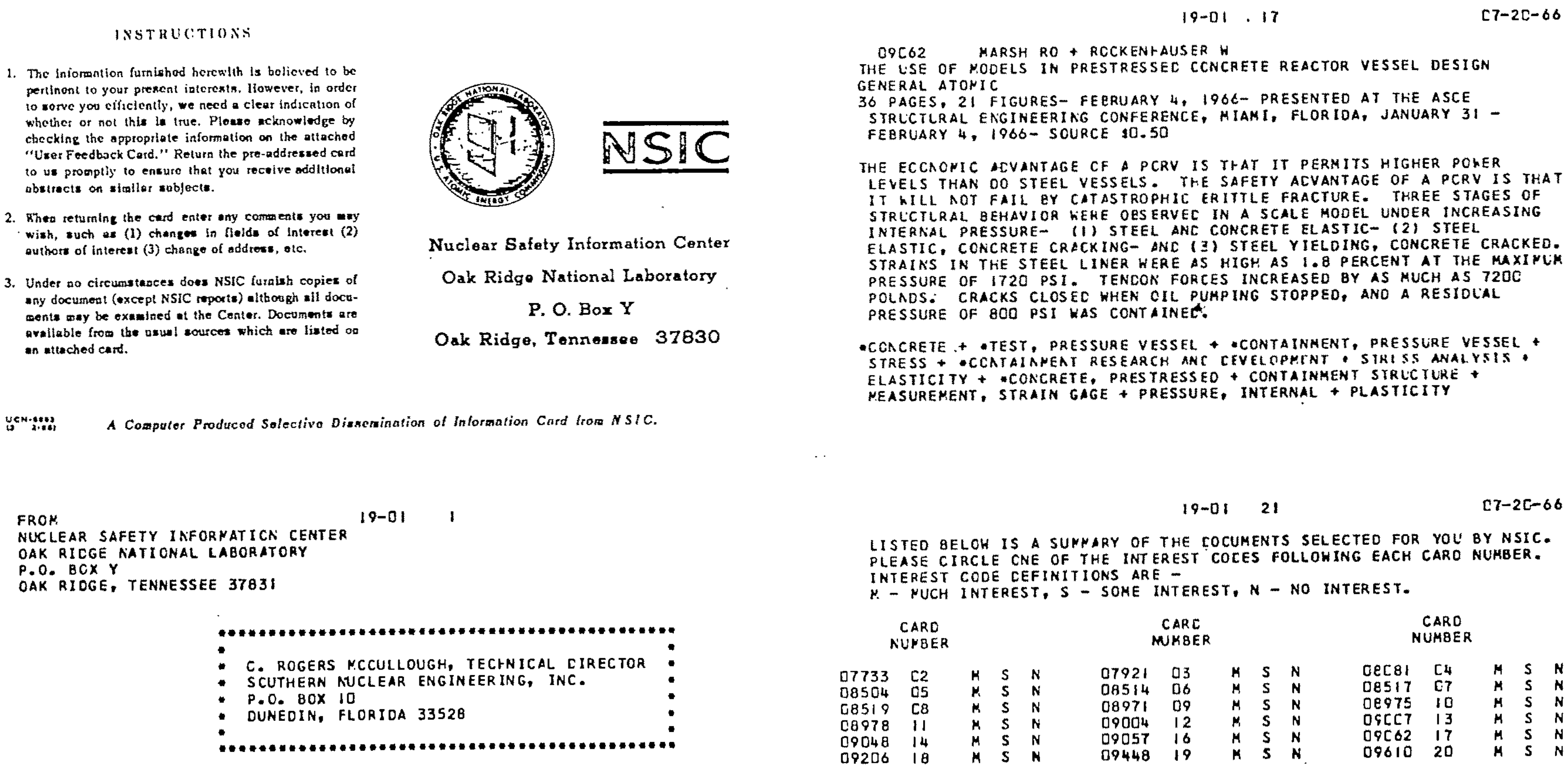
Each key word was weighted and a target score was assigned to each profile as shown in Table 1. Negative weights were permitted and provided a powerful tool to apply when developing search strategies. (One use will be discussed later in the section on profile adjustment.) A bibliographic accession on the computer tapes must equal or exceed the target score before the item is considered to meet the search parameters. An individual with varied interests was assigned more than one profile. Subject categories were then assigned to the profiles where appropriate. This limits the amount of Master Tape to be searched and conserves computer time.

NSIC's computer storage files are updated approximately every 2 weeks with the latest accessions. The SDI profiles are matched against the new material during this process. The computer scans its memory, checks each accession, and assigns the corresponding word weights to any of the key words that match those on the interest profile. When the total weight is equal to or greater than the target score, the computer prints the title, author, abstract, and key words on one of the specially prepared continuous-form 5×8 in. cards. The cards are folded in accordion style with the address card on top so that a package can be mailed to a participant rather than several individual cards. A portion of a typical SDI output is shown in Fig. 1. The first section depicts the preaddressed reverse of the card form. Also shown are an address card, a typical accession card, and a feedback card.

TABLE 1. Typical profile of an SDI recipient

Target score	Key words	Assigned weight
4	Safety system	2
	Safety review (operations, experiments)	1
	Reliability analysis	3
	Reliability system	2
	Operations report analysis	2
	Hazards analysis	2
	Inspection and compliance	2
	Administration, control, practices	3
	Operating limits/technical specification	2
	Accident probability	2
	Staffing, training, qualification	4
	Radiochemical plant safety	—1
	Radiochemical processing	—1

Categories 1, 5, 17, 18



INSTRUCTIONS

1. The information furnished herewith is believed to be pertinent to your present interests. However, in order to serve you efficiently, we need a clear indication of whether or not this is true. Please acknowledge by checking the appropriate information on the attached "User Feedback Card." Return the pre-addressed card to us promptly to ensure that you receive additional abstracts on similar subjects.
2. When returning the card enter any comments you may wish, such as (1) changes in fields of interest (2) authors of interest (3) change of address, etc.
3. Under no circumstances does NSIC furnish copies of any document (except NSIC reports) although all documents may be examined at the Center. Documents are available from the usual sources which are listed on an attached card.

NSIC

Nuclear Safety Information Center

Oak Ridge National Laboratory

P. O. Box Y

Oak Ridge, Tennessee 37830

UCN-8883
(3 2-66)

A Computer Produced Selective Dissemination of Information Card from NSIC.

```
                                 19-01  . 17                 07-20-66

  09C62      MARSH RO + ROCKENHAUSER W
THE USE OF MODELS IN PRESTRESSED CONCRETE REACTOR VESSEL DESIGN
GENERAL ATOMIC
36 PAGES, 21 FIGURES- FEBRUARY 4, 1966- PRESENTED AT THE ASCE
 STRUCTURAL ENGINEERING CONFERENCE, MIAMI, FLORIDA, JANUARY 31 -
 FEBRUARY 4, 1966- SOURCE $0.50

THE ECONOMIC ADVANTAGE OF A PCRV IS THAT IT PERMITS HIGHER POWER
 LEVELS THAN DO STEEL VESSELS.  THE SAFETY ADVANTAGE OF A PCRV IS THAT
 IT WILL NOT FAIL BY CATASTROPHIC BRITTLE FRACTURE.  THREE STAGES OF
 STRUCTURAL BEHAVIOR WERE OBSERVED IN A SCALE MODEL UNDER INCREASING
 INTERNAL PRESSURE-  (1) STEEL AND CONCRETE ELASTIC- (2) STEEL
 ELASTIC, CONCRETE CRACKING- AND (3) STEEL YIELDING, CONCRETE CRACKED.
 STRAINS IN THE STEEL LINER WERE AS HIGH AS 1.8 PERCENT AT THE MAXIMUM
 PRESSURE OF 1720 PSI.  TENDON FORCES INCREASED BY AS MUCH AS 7200
 POUNDS.  CRACKS CLOSED WHEN OIL PUMPING STOPPED, AND A RESIDUAL
 PRESSURE OF 800 PSI WAS CONTAINED.

*CONCRETE + *TEST, PRESSURE VESSEL + *CONTAINMENT, PRESSURE VESSEL +
 STRESS + *CONTAINMENT RESEARCH AND DEVELOPMENT + STRESS ANALYSIS +
 ELASTICITY + *CONCRETE, PRESTRESSED + CONTAINMENT STRUCTURE +
 MEASUREMENT, STRAIN GAGE + PRESSURE, INTERNAL + PLASTICITY
```

```
FROM                                  19-01      1
NUCLEAR SAFETY INFORMATION CENTER
OAK RIDGE NATIONAL LABORATORY
P.O. BOX Y
OAK RIDGE, TENNESSEE 37831

            *********************************************
            *                                           *
            *  C. ROGERS MCCULLOUGH, TECHNICAL DIRECTOR *
            *  SOUTHERN NUCLEAR ENGINEERING, INC.       *
            *  P.O. BOX 10                              *
            *  DUNEDIN, FLORIDA 33528                   *
            *                                           *
            *********************************************
```

```
                                 19-01    21                   07-20-66

  LISTED BELOW IS A SUMMARY OF THE DOCUMENTS SELECTED FOR YOU BY NSIC.
  PLEASE CIRCLE ONE OF THE INTEREST CODES FOLLOWING EACH CARD NUMBER.
  INTEREST CODE DEFINITIONS ARE -
  M - MUCH INTEREST, S - SOME INTEREST, N - NO INTEREST.

      CARD                       CARD                       CARD
     NUMBER                     NUMBER                     NUMBER

07733  02    M  S  N      07921  03    M  S  N      08081  04    M  S  N
08504  05    M  S  N      08514  06    M  S  N      08517  07    M  S  N
08519  08    M  S  N      08971  09    M  S  N      08975  10    M  S  N
08978  11    M  S  N      09004  12    M  S  N      09007  13    M  S  N
09048  14    M  S  N      09057  16    M  S  N      09062  17    M  S  N
09206  18    M  S  N      09448  19    M  S  N      09610  20    M  S  N
```

FIG. 1. Portions of an SDI Profile Output

The special card stock is preprinted for the Center at a cost of about 1¢ per card. The computer sorting and printing costs and mailing costs increase the expense per card to about 5¢.

• SDI Feedback and Adjustment

Feedback from the SDI recipients makes possible adjustments to individual profiles to reduce the quantity of citations of no interest. The overall feedback response has been very good; in fact, cards have been returned by over 80% of the participants. When analysis of the returns from a participant indicates that relevancy can be substantially improved, his profile is reviewed and efforts are made to make it more responsive to his needs. Physicist A will be used to illustrate how this was done. Initially this scientist was profiled to receive all of category 6 on "Reactor Transients, Kinetics, and Stability." Tabulation of the returns from the first feedback cards was as follows for 146 citations:

Documents of much interest	32%
Documents of some interest	17%
Documents of no interest	51%

Category 6 covers both analytical and experimental studies of reactors and critical facilities. An examination of the key words assigned to the documents that were of no interest to Physicist A revealed that many of them were concerned with experimental determinations or with critical assemblies. As a result, the profile was altered so that reports with either of the key words, "measurement, general," or "criticality safety," would not be cited. This was done by giving these two terms weights of "—1" with the target score held at "0" so that all other references in Category 6 would continue to be selected. Results of the next four feedback cards revealed some improvement in the pertinence, with a total of 56 citations breaking down as follows:

Documents of much interest	41%
Documents of some interest	25%
Documents of no interest	34%

Since there was room for further improvement, the next step taken was to communicate directly with Physicist A for more insight into his actual interest. This disclosed that he wanted all references in Category 6 that entail the use of reactor transient parameters but that he was not interested in the determination of these parameters. This more explicitly confirmed our analysis of the early feedback cards and suggested further profile alterations. With this information and further study of the key words assigned to the irrelevant documents, a longer list of key words was developed for negation. The following were allotted weights of "—1" and assigned to the group of two terms already designated:

Analytical model
Critical experiment
Doppler effect
Flooding coefficient
Metal water reaction
Shock wave
Temperature coefficient
Void coefficient

After this change was made, three additional feedback cards from Physicist A were reviewed. A total of 20 references was cited with the following results:

Documents of much interest	65%
Documents of some interest	25%
Documents of no interest	10%

This example illustrates clearly how individual profiles can be adjusted to close in on the user's actual needs. Further adjustments of this particular profile are probably not warranted since it has reached a range of "no interest" below which it is difficult to make significant progress. Indeed, we feel that it is probably undesirable to reduce the irrelevant figure below 10%. To do this would increase the likelihood that some relevant documents would not be cited at all. Therefore, we have concluded that an SDI profile is functioning satisfactorily if the number of irrelevant items cited falls in the range 10 to 20%.

• Quarterly Bibliography

The first computer output that NSIC distributed was the quarterly indexed bibliography. It was issued in April 1965 and contained the first 670 references that were stored on the computer tapes. The references were sorted into the 19 categories listed below into which the subject of nuclear safety has been divided.

1. General safety criteria
2. Siting of nuclear facilities
3. Transportation and handling of radioactive materials
4. Aerospace safety
5. Accident analysis
6. Reactor transients, kinetics, and stability
7. Fission product release, transport, and removal
8. Sources of energy release under accident conditions
9. Nuclear instrumentation, control, and safety systems
10. Electrical power systems
11. Containment of nuclear facilities
12. Plant safety features
13. Radiochemical plant safety
14. Radionuclide release and movement in the environment
15. Environmental surveys, monitoring and radiation exposure of man
16. Meteorological considerations
17. Operational safety and experience
18. Safety analysis and design reports
19. Bibliographies

When the references are indexed, each is assigned to at least one of the categories, but it may be assigned to as many as three, if appropriate. In the latter case, the accession information is printed in full in each of the categories rather than being cross referenced. The information for each accession is the same as that displayed on the typical SDI accession card in Fig. 1.

The bibliography program prints the categorized accessions, key words, and personal author indexes on 14 x 17 in. computer sheets. These sheets are taken two at a time and photographically reduced by about 40% to give 8 x 11 in. negatives. The pages for the bibliography are then printed from these negatives. Since the pages are numbered by the computer, and the title page, foreword, and distribution, etc., are printed by the computer, no editorial or graphic arts work is required in the process. The last issue of the quarterly bibliography contained about 300 pages and 1,000 references. Over 1,000 copies of each issue of the bibliography are distributed by NSIC.

• Computer Volumes and Running Times

At present there are over 10,700 items on the Master Tape, and the vocabulary contains almost 1,600 key words (*2*). It takes around 12 min of IBM 7090 time to enter 140 items on the Master Tape. Each key word assigned to an item is checked against the vocabulary authority to make sure that it is an authorized key word. New key words must be so designated in order to be added to the authority.

SDI. The number of scientists receiving SDI service has grown from 50 to over 800 since the program started in October 1965. A scientist may have more than one profile of interest, and these profiles now exceed 1,000. All together, the profiles query 183 authors, 93 corporate authors, 2,375 categories of information, and 4,821 key words. One scientist has 9 profiles querying 20 categories and 310 key words.

All question information is held in the computer core, and each item from the Master Tape is worked against it. One item may drop for several participants, and a subsequent sort is used to pull together all drops for one participant. At present it is possible to process some 200 profiles with only one reading of the Master Tape. The tape is then rewound, and the next group of profiles is processed similarly.

Total computer time for an SDI run that dropped 375 items for 165 profiles was around 15 min. Of this 15 min, the actual questioning and selecting required only 2 min. The rest of the time was taken up in sorting, formatting for output, etc.

Bibliography. Ten quarterly bibliographies have been issued, each containing 600 to 1,000 references. The last issue required 24 min of IBM-7090 time for sorting into the NSIC categories and preparation of the key word and personal author indexes.

• Principal NSIC Services

A variety of informational services is offered by NSIC to the nuclear community. The principal ones are summarized below. The computer outputs discussed in more detail above are included here for completeness. The *Nuclear Safety* Journal is available by subscription only. All other NSIC services, including state-of-the-art reports, may be obtained from the Center without charge by persons who are active in the nuclear field. Those who do not fall in this category may purchase such reports for a nominal sum from the Clearinghouse for Federal Scientific and Technical Information, Springfield, Virginia.

State-of-the-Art Reports. These reports provide a mechanism for the individual staff members to analyze and evaluate the experimental and theoretical data developed in their particular subject area. They are very comprehensive and require several man-months of technical effort to produce. Subjects that have been or are currently being covered include: (1) iodine monitoring practices (*3*), (2) iodine behavior in reactor containment systems (*4*), (3) reactor containment practices (*5*), (4) reactor secondary shutdown systems (*6*) (5) nuclear safety research and development projects (*7*), (6) U.S. nuclear standards (*8, 9*), (7) air-cleaning systems (*10*), (8) reactor pressure vessel integrity (*11*), (9) reactor operating experiences (*12*), (10) international nuclear standards (*13*), (11) height of rise of effluents, and (12) tritium behavior.

Nuclear Safety. The technical progress review Nuclear Safety, while separately funded, is prepared by NSIC. Recent developments in nuclear safety are concisely reviewed as prevailing interest and available information warrant (*14*). The journal may be purchased by subscription from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402, for $3.00 per year.

Indexed Bibliography of Accessions. The bibliographic accessions of the Center are published quarterly. This computer output is sorted according to the 19 NSIC categories and issued with key words and personal author indexes. So far, 10 have been published (*15–24*). A bibliography of all the accessions in Category 3 has also been published (*25*).

Selective Dissemination of Information. Bibliographic citations on continuous 5 x 8 in. cards are mailed to participants on a biweekly basis. The citations are selected according to each participant's profile of interest as the computer tapes are updated. The operation of this important program is described above in some detail.

Technical Inquiries. Inquiries for nuclear information are currently being received and answered, free of charge, by telephone, letter, or personal contact at a rate of 40 per month. This is more than double the rate at which requests were received in 1965.

Counseling and Guidance. The NSIC staff is available to visitors for counseling and guidance on nuclear safety problems in its subject areas. Visits to the Center for the purpose of consulting with the technical staff or using the information storage files occur at a rate of about 12 per month.

Special Bibliographies. The computer files may be queried for special retrospective searches at any time.

The output is a bibliography appropriate to the particular search combination of key words, authors, date, and/or evaluation. The information printed in the output is the same as that for the SDI; that is, the same as that displayed on the accession card in Fig. 1.[2]

• Organizational Structure

The Nuclear Safety Information Center is funded through the office of the Assistant Director for Nuclear Safety, Dr. J. A. Lieberman, in the AEC Division of Reactor Development and Technology. The AEC Division of Technical Information collaborates in sponsorship of the Center. The Center is organized at the laboratory as a semi-autonomous group, responsible to an Assistant Laboratory Director, who is in turn responsible for all informational problems; but it is administered within the ORNL Reactor Division. At the present time the staff is composed of 20 technical specialists (most of whom are on a part-time basis), a technical editor, two information specialists, and four secretary-typists as given in Table 2. A staff scientist or engineer is assigned to each of categories 6, 7, 9, 11, 14, 15, 16, and 17 on a half-time basis. Surveillance in the other areas is on a lesser level, usually approximating a tenth time in each. In the other fraction of their time, these specialists work in an experimental or analytical group active in their particular subject area.

TABLE 2. Organization of the NSIC

Wm. B. Cottrell, Director
Joel R. Buchanan, Assistant Director
Jeanne Thomas, Secretary
H. B. Whetsel, Technical Editor
Celia Murphy, Chief Information Specialist
Shirley Hendrix, Information Specialist (part-time)
Ann Hayes, Information Specialist and Secretary
Janet Davis, Secretary
Dianne Lane, Secretary

Staff Scientists and Engineers

J. P. Blakely, Physical Chemist
J. R. Buchanan, Nuclear Engineer
Wm. B. Cottrell, Nuclear Engineer
E. N. Cramer, Nuclear Engineer
W. K. Ergen, Physicist
M. H. Fontana, Nuclear Engineer
S. D. Swisher, Meteorologist
D. G. Jacobs, Soil Chemist
G. W. Keilholtz, Physical Chemist
C. G. Lawson, Nuclear Engineer
T. F. Lomenick, Geologist
W. C. McClain, Geophysicist
H. A. McLain, Nuclear Engineer
J. G. Merkle, Mechanical Engineer
H. B. Piper, Nuclear Engineer
J. B. Ruch, Chemical Engineer
R. L. Scott, Physicist
L. B. Shappert, Nuclear Engineer
C. S. Walker, Electrical Engineer
M. L. Winton, Nuclear Engineer

The director of the information center is coordinator of ORNL's nuclear safety research and development program and editor of the technical progress review *Nuclear Safety*. The assistant director, who is responsible for the day-to-day operation of the Center, is an assistant editor of *Nuclear Safety*. The technical staff is composed entirely of senior people who have been working in their areas of speciality for a number of years and who continue to work in these areas. These scientific middlemen are the backbone of the NSIC. In fact, as it was so aptly expressed in the "Weinberg Report" in 1963, "The essence of a good technical information center is that it be operated by highly competent working scientists and engineers—people who see in the operation of the Center an opportunity to advance and deepen their own personal contact with their science and technology (*26*).

• Future Plans

Just recently, NSIC received delivery on the equipment for a computer telecommunications system, IBM-2260, which will be located at NSIC and will have the capability for (1) direct data input to the computer, (2) scanning and querying of the computer data files, and (3) direct maintenance of the computer files. Information may be displayed on cathode-ray screens for scanning and modification. The devices will be coupled to the IBM-360 at the Computing Technology Center. All these information-processing techniques represent significant improvements in NSIC's information system and will enable it to efficiently function as part of the national information system of the future that will encompass all science and technology.

References

1. BUCHANAN, J. R., and W. B. COTTRELL, *Operating Experience of the Nuclear Safety Information Center March 1963–March 1965*, USAEC Rep. ORNL–TM–1136, Oak Ridge National Laboratory, Oak Ridge, Tenn., May 17, 1965.
2. *NSIC Key-word Thesaurus*, USAEC Rep. ORNL–NSIC–35, Oak Ridge National Laboratory, Oak Ridge, Tenn. August 1967.
3. COWSER, K. E., *Current Practices in the Release and Monitoring of ^{131}I at NRTS, Hanford, Savannah River and ORNL*, USAEC Rep. ORNL–NSIC–3, Oak Ridge National Laboratory, Oak Ridge, Tenn., August 1964.
4. KEILHOLTZ, G. W., and C. J. BARTON, *Behavior of Iodine in Reactor Containment Systems*, USAEC Rep.

[2] Requests for NSIC services or additional information on these services may be addressed to: Nuclear Safety Information Center, Oak Ridge National Laboratory, P.O. Box Y, Oak Ridge, Tennessee 37830.

ORNL–NSIC–4, Oak Ridge National Laboratory, Oak Ridge, Tenn., February 1965.

5. Cottrell, W. B., and A. W. Savolainen, *U.S. Reactor Containment Technology—A Compilation of Current Practice in Analysis, Design, Construction, Test, and Operation,* Vol. I and II, USAEC Rep. ORNL–NSIC–5, Oak Ridge National Laboratory, Oak Ridge, Tenn., August 1965.

6. Walker, C. S., *Secondary Shutdown Systems of Nuclear Power Plants,* USAEC Rep. ORNL–NSIC–7, Oak Ridge National Laboratory, Oak Ridge, Tenn., January 1966.

7. Buchanan, J. R., and N. F. Cross, *Current Nuclear Safety Research and Development Projects,* USAEC Rep. ORNL–NSIC–10, Oak Ridge National Laboratory, Oak Ridge, Tenn. June 1966.

8. Cottrell, W. B., and ASA Subcommittee N6.9, *Compilation of U.S. Nuclear Standards,* 2nd ed., 1965, USAEC Rep. ORNL–NSIC–11, Oak Ridge National Laboratory, Oak Ridge, Tenn., December 1965.

9. Cottrell, W. B., Compilation of United States Nuclear Standards, 3rd ed., USAEC Rep. ORNL–NSIC–30, Oak Ridge National Laboratory, Oak Ridge, Tenn., December 1966.

10. Keilholtz, G. W., *Air Cleaning Systems—Fundamental Processes, Testing, and Nuclear Applications,* USAEC Rep. ORNL–NSIC–13, Oak Ridge National Laboratory, Oak Ridge, Tenn., September 1966.

11. Miller, E. C., *The Integrity of Reactor Pressure Vessels,* USAEC Rep. ORNL–NSIC–15, Oak Ridge National Laboratory, Oak Ridge, Tenn., May 1966.

12. U.S. Atomic Energy Commission Division of Operational Safety, *Abnormal Reactor Operating Experiences,* USAEC Rep. ORNL–NSIC–17, Oak Ridge National Laboratory, Oak Ridge, Tenn., August 1966.

13. Cottrell, W. B., and ASA Subcommittee N6.9, *Compilation of National and International Nuclear Standards* (Excluding U.S. Activities), 2nd ed., 1966. USAEC Rep. ORNL–NSIC–18, Oak Ridge National Laboratory, Oak Ridge, Tenn., June 1966.

14. *Index to Nuclear Safety A Technical Progress Review By Chronology,* Permuted Title, and Author Vol. 1, No. 1 Through Vol. 7, No. 4, USAEC Report ORNL–NSIC–31, Oak Ridge National Laboratory, Oak Ridge, Tenn., January 1967.

15. Nuclear Safety Information Center Staff, *Indexed Bibliography of Current Nuclear Safety Literature-1,* USAEC Rep. ORNL–NSIC–8, Oak Ridge National Laboratory, Oak Ridge, Tenn., April 1965.

16. Nuclear Safety Information Center Staff, *Indexed Bibliography of Current Nuclear Safety Literature-2,* USAEC Rep. ORNL–NSIC–9, Oak Ridge National Laboratory, Oak Ridge, Tenn., August 1965.

17. Nuclear Safety Information Center Staff, *Indexed Bibliography of Current Nuclear Safety Literature-3,* USAEC Rep. ORNL–NSIC–12, Oak Ridge National Laboratory, Oak Ridge, Tenn., November 1965.

18. Nuclear Safety Information Center Staff, *Indexed Bibliography of Current Nuclear Safety Literature-4,* USAEC Rep. ORNL–NSIC–14, Oak Ridge National Laboratory, Oak Ridge, Tenn., March 1966.

19. Nuclear Safety Information Center Staff, *Indexed Bibliography of Current Nuclear Safety Literature-5,* USAEC Rep. ORNL–NSIC–16, Oak Ridge National Laboratory, Oak Ridge, Tenn., June 1966.

20. Nuclear Safety Information Center Staff, *Indexed Bibliography of Current Nuclear Safety Literature-6,* USAEC Rep. ORNL–NSIC–19, Oak Ridge National Laboratory, Oak Ridge, Tenn., September 1966.

21. Nuclear Safety Information Center Staff, Indexed Bibliography of Current Nuclear Safety Literature-7, USAEC Rep. ORNL–NSIC–20, Oak Ridge National Laboratory, Oak Ridge, Tenn., November 1966.

22. Nuclear Safety Information Center Staff, Indexed Bibliography of Current Nuclear Safety Literature-8, USAEC Rep. ORNL–NSIC–32, Oak Ridge National Laboratory, Oak Ridge, Tenn., March 1967.

23. Nuclear Safety Information Center Staff, *Indexed Bibliography of Current Nuclear Safety Literature-9,* USAEC Rep. ORNL–NSIC–34, Oak Ridge National Laboratory, Oak Ridge, Tenn., 1967.

24. Nuclear Safety Information Center Staff, *Indexed Bibliography of Current Nuclear Safety Literature-10,* USAEC Rep. ORNL–NSIC–36, Oak Ridge National Laboratory, Oak Ridge, Tenn., August 1967.

25. Shappert, L. B., and R. S. Burns, *Indexed Bibliography on Transportation and Handling of Radioactive Materials,* USAEC Rep., ORNL–NSIC–33, Oak Ridge National Laboratory, Oak Ridge, Tenn., June 1967.

26. The President's Science Advisory Committee, *Science, Government, and Information,* U.S. Government Printing Office, Washington, D.C., 1963.

Opinion Paper

The Interpretation of SDI Data

Although a large number of Selective Dissemination of Information (SDI) Systems have been planned, implemented, and tested over the past few years, insufficient attention has been given to the collection and interpretation of important data needed for evaluation. We describe some of the defects common to almost all of the reported systems, single out one recent report for detailed discussion and argue in favor of collection and correct interpretation of data on one important and frequently overlooked evaluation factor.

T. R. SAVAGE

Control Data Corporation

• Introduction

During the past 5 years numerous reports of Selective Dissemination (SDI) Systems, under development, in operation proposed for implementation, being tested, etc., have appeared in the literature. (*1–55*) It seems to me that almost all of these reports have some serious deficiencies. Most seriously, they provide little or no ground for comparison with other systems.[2]

As a participant in the development of the first SDI Systems (from 1959–1963), I became sensitive to problems of SDI evaluation and have continued to be concerned with the proper application of evaluative measures. I would like, therefore, to discuss the general problem of SDI System evaluation, using as a vehicle for the discussion an analysis of the recent paper by C. R. Sage on the evaluation of the Ames SDI System. (*56*) I chose the Sage report for particular attention first, because, it is the most recent, and second, because it provides in considerable detail, the illusion of having performed significant analyses and evaluation of the SDI data.

• *General SDI Features*

Before examining the Ames System in detail, some basic clarifications of the important features of all SDI Systems need to be made. The basic universe of discourse (or population) that is pertinent to SDI Systems is, as I have urged elsewhere, (*57*) not simply users (U) or documents (D), but their product, (UD). Another important, but frequently overlooked item of information is the percentage of UD that results in notices to the users. This might be called the "selective reaction" of the system. Note that UD represents all possible notices that might be sent, that is, everybody gets everything; and the purpose of SDI is to reduce this total.

Although it has become commonplace, since the results of the Cleverdon work have become widely known, (*58–60*) to talk in terms of relevance and recall as evaluation measures, these can be seen to be not really appropriate when one realizes: first, that the object of our study is UD rather than some subsets of it; and second, that evaluation measures should be viewed as statistical estimates of system performance. To revert to an early terminology used elsewhere, (*10*) we can label the four boxes of the familiar 2 x 2 matrix (shown in Table 1) as: 1-Hits 2-Trash; 3-Miss; 4-Pass. Trash and miss can be easily recognized as analogs of our old statistical friends Type 1 and Type 2 error. A reasonably satisfactory first step toward providing some evaluation measure is simply to compute the reciprocal of the sum of miss and trash. This provides higher scores for systems with less error and vice versa. This measure is comparative (like a scratch test for hardness) and not metric. The metricizing of evaluation measures is a separate and complicated problem that will not be treated here.

What is yet unknown about any information system (SDI or otherwise) is, in any precise form, the relative

[1] The extensiveness of the SDI literature was not apparent to me when work began on this paper. I started searching the literature as the work proceeded and the list of References 41-55 is the result. Although the list is not exhaustive, it should be helpful since the only survey of SDI literature is the now outdated work of Hensley (Reference 214).

[2] Two exceptions to this indictment are the systems described in References 9 and 44.

Table 1. Two X two matrix showing subsets of UD

	Accepted by user	Not accepted by user
System selected	1. Hits	2. Trash
System not selected	3. Miss	4. Pass

values that users (either individually or collectively) assign to these four factors. Users that are starved for information will tend to value hits and not be disturbed by large amounts of miss or trash. Users flooded with information will value pass and strongly disvalue trash, etc.

• The Ames System

In the concluding paragraph of Sage's paper, he states:

> The relative percentage of interest computed as illustrated in Section IV of this article may tend to appear low compared with other SDI Systems. However, our only criteria for measuring any degree of worthwhile service is through the reactions and comments of our users.

Apparently he regards the second sentence as somehow an explanation for the first. Of course, it is not. To draw an analogy (as Taube did) (*61*) between our field and the practice of medicine, Sage's statement could be recast as:

> It's true that my patients die as a result of my treatment. However, the only criterion for measuring any degree of worthwhile (medical) service is through the reactions of patients.

It is unfortunate that Sage has taken this rather disingenuous attitude toward his system, since (1) the paper itself provides much of evidence to account for the Ames System's deficiencies and (2) despite the detailed analyses of the notifications supplied by the Ames System, there apparently is no method used to determine miss.[3]

In the Ames evaluation test there were two separate document populations under study: those from Nuclear Science Abstracts (*NSA*) and those from the Science Citation Index (*SCI*). Both were run against 21 User Profiles. For NSA, UD was 532,938. The number of notices generated was 11,226,[4] for a selective reaction of .02. For SCI, US was 1,456,308. The number of notices generated was 6,038, for a selective reaction of .004. This difference in selective reaction by a factor of 5 was occasioned by a difference in threshold value of only a factor of about 1.7. When one looks for the source of this discrepancy, it is apparently found in the average depth of indexing of the two document populations. NSA had a depth of 43.1 terms and SCI a depth of 11.2 terms. Although I have been unable to determine from Sage's paper how the matching function and threshold calculation actually work, it would appear that they simply produce more notices as an almost direct function of the number of words in the document profile. Noticing this problem at all, however, requires attention to the selective reaction of the system.

In his treatment of "impartial" responses, Sage excludes these in his evaluation calculations. This is not quite fair. Strictly these should be counted as trash. Recomputing the results on this basis gives the results shown in Table 2.

The hit ratios, or percentage of sent notices, the users judged of interest, are 30% and 31%, respectively. These are very low indeed.

Since the major differences between the Ames SDI System and others with which I am familiar are the use of the significance values, the threshold calculations, and the feedback adjustments, these features may be the source of some of the difficulty. The data for the whole of 1965 (Sage's Section IV) show 81 users matched against 177, 180 documents for a UD of 14,351,580; 54,018 notices were generated for a selective reaction of .004. Sufficient data are not provided to calculate hit ratios, but one could expect the results to be similar to those shown above.

For comparison purposes, let us look at the data for three other SDI Systems that have been developed and tested.[5] These are shown in Table 3. The Ames System fares rather badly in all categories.

Strictly, of course, the numbers in both Table 2 and Table 3 should be shown as statistical estimates with variances, levels of confidence, and preferably confidence limits indicated for each value. On intuitive grounds, we usually assume that miss and pass are somehow more "estimates" than hits and trash, because the former are measured indirectly, while the latter are directly computed. This view is mistaken for at least three reasons.

Table 2. Ames SDI system data

	NSA		SCI	
	Number	% U D	Number	% U D
Hits	3,387	.006	1,896	.001
Trash	7,839	.014	4,142	.003
Miss	*	*	*	*
Pass	521,706	.98	1,450,270	.996

* In the Ames System miss is unknown and has been lumped with pass.

[3] Frequently the argument is used that we cannot adequately determine miss without an exhaustive examination of the entire document collection. This is simply a mistaken notion of scientific procedure. Random sampling as used in Reference 9 is a perfectly acceptable and straight-forward method for determining miss.

[4] These numbers as well as those shown in Table 2, were computed from the average data supplied by Sage.

[5] In Table 3, "SDI-1" is the system described by Reference 9, "KRAFT" is the system described in Reference 24, and "SDI-2" is the system described in Reference 3. The data in this report were supplemented with unpublished data available to the author.

TABLE 3.

	SDI-1		KRAFT		SDI-2	
	Number	% U D	Number	% U D	Number	% U D
Hits	415	.06	17,350	.009	14,629	.023
Hit Ratio		.41*		.66*		.68*
Trash	597	.09	8,950	.005	6,641	.01
Miss	385	.06	†	†	227,308	.36
Pass	5,273	.79	1,833,700	.986	370,972	.607
S R		.13		.014		.036

* These are not %'s of UD, of course, but % of the sum of Hits and Trash.

† Again, Miss was unknown and lumped with Pass.

In the first place, the basis on which we accept a value as correct is not how it is measured, but rather, the confidence, statistically estimated, with which we can assume the value is the true one. Secondly, since hits and trash are defined in terms of individual users responses, any aggregrate numbers for whole systems are subject to the same statistical treatment which we apply to any other sampling of a population. (Remember that the selective reaction is in fact a method, and hopefully a biased one, of sampling UD.) Finally, the main purpose of obtaining evaluation data at all is to use them as predictors of system operation. All predictors are, of course, estimates of future performance.

• The Problems of Miss

Aside from the fact that the Ames System operates at a rather low level of efficiency,[6] its failure to provide for an estimate of miss leaves us in a rather poor position for recommending any effective means of correction. The simplest method of estimating miss is to send some notices randomly to the users, determine the hit ratio of the random notices, use this percentage as an estimate of the good user-document combinations, multiply all of UD by this percentage to get an estimate of the actual number of good user document combinations, and subtract from this number the good notices actually sent (hits). The remainder is then the estimate of miss.

The reasons for providing estimates of miss are not widely appreciated and perhaps should be enumerated here. In the first place, the common use of hit ratios is deceptive if used alone as a measure of performance. In most SDI Systems it is a relatively simple matter to obtain a high hit ratio by requiring a higher proportion of match between the user and document profiles. This reduces the number of notices and sends only those with a high probability of being judged relevant. But, also, (because of the reduced notices) this increases miss. Secondly, without an estimate of miss; we have no method of determining the a priori match between our user and document populations. If, for example, we matched MEDLARS documents to AEC Scientists, we may get a very low hit ratio, but miss may also be very low, indicating that the users and documents just don't match each other. In such a case, we are better advised to select another document (or user) population before trying to modify our system. In the SDI-2 case, noted above, miss was very high, indicating a high a priori match between user and document populations, so that the high hit ratio is shown, in part at least, to be an artifact of the user's a priori interest in the documents. One system reports a hit ratio of .91.[7] This is very high indeed. Without, however, an estimate of miss, we don't know whether to applaud the SDI System or the acquisitions department.

There is some subtle, and, as yet, unknown relation among selective reaction, the a priori match between users and documents, the ability of users to absorb notices (or documents), and the number of documents the system chooses to process. In SDI System operation, there are two factors of importance: (1) the product of the reading (or scanning) speeds of the users and the time the users have available for SDI and (2) the amount of time the user is willing to wait without receiving some (unknown amount of) relevant notices before abandoning SDI altogether. These numbers are not easy to obtain, but serious attempts should be made to estimate them, lest the SDI System overburden the user with paper or starve him of notices. Estimates of miss can help in this task.

For example, if the system selective reaction is high and the a priori match is high, the system must restrict the number of documents it processes in order not to flood the users with notifications. On the other hand, if both the selective reaction and the a priori match are low the system must process many, many documents in order to generate enough notices to give any service at all.[8] For simple economic reasons high selective reaction and high a priori match are to be preferred. The first of these is a system processing problem, and the second an acquisitions problem. Without knowing miss, we don't know which.

The third reason for emphasizing the collection of data on miss is that we force on the system operators the necessity for accounting for system error, and thereby have grounds for recommending improvement.

Finally, obtaining estimates of miss and trash gives us, for the first time, data for actual comparisons among alternative systems.

[6] Curiously, the system operating costs of Ames fare badly when compared with SDI-2. The SDI-2 processing rate was, approximately, in minutes .0018 × UD, while the Ames processing rate is, approximately, in minutes, .0076 × UD. Ames runs on an IBM 7074/1401. SDI-2 was run with a Fortran program on an IBM 704.

[7] This is the Douglas System as reported in Reference 39.

[8] This apparently is exactly the difficulty encountered at Ames.

Conclusion

SDI remains the least expensive, most effective and most easily evaluated system to use as a base of information services. Unless, however, the system is evaluated correctly and the results of the evaluation are judiciously used to modify and improve performance, the advantages of SDI quickly disappear. SDI Systems, although simple in concept, are, like any dynamic system interacting with humans, complex in actual operations.

References

1. Sowarby, A. J., The Selective Dissemination of Information System-Present Operations and Future Applications, *in* C. F. Balz, Ed., *Library Seminar—October 19-20, 1960,* IBM Federal Systems Division, Space Guidance Center, Owego, N.Y., February 28, 1961, pp. 14–40.
2. Luhn, H. P., Selective Dissemination of New Scientific Information With The Aid of Electronic Processing Equipment, *American Documentation,* 12 (No. 2): 131–138 (1961).
3. Brandenberg, W., et al., Selective Dissemination of Information, SDI—2 System, *IBM Advanced Systems Development Division,* Rep. No. 17–031, (1961).
4. Veyette, J. H., Jr., Information Retrieval, *American Behavioral Scientist,* (No. 10): 15–20 (1961).
5. Benjamin, R., et al., *Selective Dissemination of Information,* IBM Corporation, Federal Systems Division, Owego, N.Y., 1961, Rep. No., 61–510–7.
6. Resnick, A., Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents, *Science,* 134:1004–1006 (1961).
7. Dennis, B. K., New Concepts in Technical Information Services; *Proceedings of The Engineering Information Symposium, Engineers Joint Council* (1962).
8. Koriagin, G. W., and L. R. Bunnow, Mechanized Information Retrieval System for Douglas Aircraft Company, Inc., Status Report, Rep. No. SM–39167, Douglas Aircraft Co., Inc., (1962).
9. Hensley, C. B., et. al., Selective Dissemination of Informations—A New Approach to Effective Communication, *IRE Transactions On Engineering Management,* EM–9 (No. 2): pp. 55–65 (June 1962).
10. Resnick, A., Comparative Effect of Different Education Levels on Indexing in a Selective Dissemination System, Rep. No. 17–092, *IBM Advanced Systems Development Division,* (1962).
11. Carroll, K. D., and R. K. Summit, MATICO: Machine Applications To Technical Information Center Operations, Lockheed Aircraft Corporation, Rep. No. 5–13–62–1, Missiles and Space Systems Division, Sunnyvale, California, September 1962.
12. Koriagin, G. W., Library Information Retrieval Program," *Journal Of Chemical Documentation,* 2 (No. 4) 242–248 (1962).
13. Sowarby, A. J., SDI-3 For The IBM 1401 Data Processing System, File No. 10.3.004, IBM Corporation, 1401 General Program Library, 1962.
14. Timberlake, W. D., An Information Retrieval and Dissemination System, Rep. No. TR00.930 IBM DSD Development Laboratory, Poughkeepsie, N.Y., 1962.
15. Benjamin, R., et al., Selective Dissemination Information (SDI) Share Distribution No. 1372. Space Guidance Center, Owego, N.Y., 1962.
16. Koriagin, G. W., Experience In Man and Machine Relationships In Library Mechanization," Paper No. 1495. Douglas Missiles and Space Systems Division, Engineering, 1962. (A condensed version with the same title can be found in *American Documentation,* 15, 227–229, 1964.
17. Kraft, D. H., Data Processing Equipment For Library Use In Clerical Tasks and Dissemination of Information, *Illinois Libraries,* 44:587–592 (1962).
18. Tritschler, R. J., A Computer-Integrated System For Centralized Information Dissemination, Storage and Retrieval," *ASLIB Proccedings,* 14:473–503 (1962).
19. Wadding, R. V., and R. H. Stanwood, *The MERGE System Package,* Share Distribution No. 1465, IBM Corporation, Space Guidance Center, Owego, N.Y., 1963.
20. Resnick, A., and C. B. Hensley, The Use of Diary and Interview Techniques in Evaluating a System for Disseminating Technical Information, *American Documentation,* 14:109–116 (1963).
21. Hensley, C. B., Selective Dissemination Of Information (SDI): State Of The Art In May 1963, *Spring Joint Computer Conference,* (1963).
22. Balz, C. F., and R. S. Stanwood, Literature Dissemination and Retrieval Using the MERGE System, *Automation and Scientific Communication,* Short Papers —Part 1, 61–62 (1963).
23. Freeman, R. R., Automatic Retrieval and Selective Dissemination Of References From Chemical Titles: Improving The Selection Process," *Automation and Scientific Communication,* Short Papers—Part 2, pp. 213–214 (1963).
24. Kraft, D. H., An Operational Selective Dissemination of Information/SDI System for Technical and Non-Technical Personnel Using Automatic Indexing Techniques, *Automation and Scientific Communication,* Short Papers—Part 1, 69–70 (1963).
25. Hilf, J. W., Matching Descriptions in A Selective Dissemination System, *Automation and Scientific Communication,* Short Papers—Part 1, 65–66 (1963).
26. Resnick, A., "Progress Report On IBM's Selective Dissemination Of Information—SDI-4 System—IBM 7090–1401 Data Processing System," *Automation and Scientific Communication,* Short Papers—Part 2, 329–330 (1963).
27. Resnick, A., Educational Requirements For Indexers In A Selective Dissemination System, *Automation and Scientific Communication,* Short Papers—Part 2, 163–164 (1963).
28. White, H. S., The IBM DSD Technical Information Center—A Total Operating System Approach Combining Traditional Library Features and Mechanized Computer Processing, *Automation and Scientific Communication,* Short Papers—Part 2, 287–288 (1963).

29. White, H. S., DSD Technical Information Center—A Total Approach To Library Mechanization, *Automation and Scientific Communication*, Proceedings—Part 3, pp. 443–449 (1963).

30. Ofer, K. D., SIDAR: Selective Information Dissemination And Retrieval, *Journal Of Chemical Documentation*, 4:54–55 (1964).

31. Barnes, A. B., et al., SDI-5 An Advanced System For Selective Dissemination Of Information, *ACM, Proceedings Of The 19th National Conference*, No. L2.2 (1964).

32. Anderson, R. R., et. al., The Stability Of Dynamic Feedback In An SDI System," *Proceedings Of The 19th International Guide Meeting* (1964).

33. Merritt, C. A., The User and The Technical Information Center, *Proceedings of the American Documentation Institute*, 1:155–158 (1964).

34. Resnick, A., The Information Explosion and the User's Need for Hard Copy," *Proceedings Of The American Documentation Institute*, 1:315–318 (1964).

35. Sage, C. R., and D. R. Fitzuater, Operational Results of an Adaptive SDI System, *Proceedings of the 19th International Guide Meeting* (1964).

36. Sage, C. R., et. al., Ames Laboratory SDI Reference Manual, *USAEC*, Report No. IS-940 (1964).

37. Wilson, R. A., (Review Of (31)), *Computing Reviews*, 6, (No. 1) (January–February 1965).

38. Magnino, J. J., Jr., *Normal Test Techniques* Report No. ITIRC-002. IBM Technical Information Retrieval Center, Thomas J. Watson Research Center, Yorktown Heights, N.Y., (1965.)

39. Young, E. J., and A. S. Williams, *Historical Development And Present Status—Douglas Aircraft Company Computerized Library Program,"* Paper No. 3453, Douglas Missiles and Space Systems Division, Douglas (May 1965).

40. Sage, C. R., et. al., "Adaptive Information Dissemination," *American Documentation*, 16:185–200 (1965).

41. Merritt, C. A., and P. J. Nelson, *The Engineer-Scientist An Information*, Report No. ITIRC-003, IBM Technical Information Retrieval Center, Thomas J. Watson Research Center, Yorktown Heights, N.Y., 1965.

42. Balz, C. F., and R. H. Stanwood, MERGE: A Current-Awareness And Retrospective Searching System For Technical Documents, *in Colloquium On Technical Preconditions for Retrieval Center Operations*, Spartan Books, 1965, pp. 61–70.

43. Computer Usage Co., Inc., *Selective Dissemination Of Information*, AD-617086, CFSTI, 1965.

44. Sprague, R. A., Jr., A Comparison Of Systems For Selectively Disseminating Information, *Bureau Business Research, Graduate School Of Business, Indiana University*, Rep. #38, (1965).

45. Ebersole, J. L., An Operating Model of a National Information System, *American Documentation*, 17:33–40 (1966).

46. Shank, R., "Letter to the Editor," *American Documentation*, 17:45 (1966).

47. Becker, J., (Review of (41)), *Computing Reviews*, 7 (No. 3): (1966).

48. Koller, H. R., (Review of (45)), *Computing Reviews*, 7 (No. 3): (1966).

49. Tritschler, R. J., (Review of (10)), *Computing Reviews*, 7 (No. 3): (May-June 1966).

51. Holzbaur, F. W., and E. K. Parker, Selective Information Dissemination by Computer—Based Subject Category Assignment, *Proceedings Of The American Documentation Institute*, 3:35–42 (1966).

51. Magnino, J. J., Jr., IBM Technical Retrieval Center Progress and Plans," *Proceedings Of The American Documentation Institute*, 3:467–480 (1966).

52. Sage, C. R., The Utilization Of A National Or Regional Current Awareness System For Retrospective Information Retrieval, *Proceedings Of The American Documentation Institute*, 3:107–116 (1966).

53. Smith, F. R., and S. O. Jones, Five Years In Focus—The Douglas Aircraft Company Mechanized Information System, *Proceedings Of The American Documentation Institute*, 3:185–192 (1966).

54. Stanwood, S. H., Some Observations On User Response To An SDI System, *Proceedings Of The American Documentation Institute*, 3:235–244 (1966).

55. Sage, C. R., et al., *Ames Laboratory Selective Dissemination of Information General Information Preliminary Report, Iowa State University*, Ames, 1966.

56. Sage, C. R., Comprehensive Dissemination of Current Literature, *American Documentation*, 17:155–177 (1966).

57. Savage, T. R., Users versus Documents, *American Documentation*, 17:141 (1966).

58. Cleverdon, C., The Evaluation of Systems Used in Information Retrieval, *Proceedings Of The International Conference On Scientific Information, National Academy of Sciences*, 1:687–698 (1959).

59. Cleverdon, C., *ASLIB Cranfield Research Project: Report on the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems*, College of Aeronautics, Cranfield, England (1960).

60. Cleverdon, C., *Interim Report on the Test Programme of an Investigation into the Comparative Efficiency of Indexing Systems*, The College of Aeronautics, Cranfield, England, 1960.

61. Taube, M., The Coming Of Age Of Information Technology, *Bulletin of the Medical Library Association*, 52:120–127 (1964). (Reprinted with some modifications, including dropping the first "of" in the title, in *The Coming Age Of Information Technology, Studies In Coordinate Indexing*, VI: 1–10 [1965].

Brief Communications

Suggestions on How to Read Experimental Material in Information Science

1. Understanding experimental work in information science takes active participation on the part of the reader. Read the purpose, proposed methods, and results or conclusions in every report first. Then read the whole actively, not passively, and with a critical eye. The more material read, the easier it becomes to understand because much experimental work falls into general patterns, and is based on a few fundamental ideas which are not too difficult to recognize.

Watch for sweeping statements. These are usually in the introductory sections, the conclusions, or in projections into the future or into other fields. Adopt a "show me" attitude regarding every statement of fact.

2. Do not be scared of big words, mathematical formulas, tables of figures, or fancy diagrams. Look up undefined words in an unabridged dictionary. If they are not there and they are not explained in the text and you cannot make sense of the report without knowing what they mean, do not waste your time reading it further. The author is not trying to communicate, but has some other motive in publishing.

Skip the mathematical formulas and look at the concepts or conclusions which they represent. Since mathematics is a shorthand language for what is explained in the text, very rarely are formulas needed to understand the work.

Add up the amounts in tables or figures and, if you can, recalculate them. These often come to something like 154% or to 31 when the author says he used 25 examples. The odd figures mean he has done something to or with them. This may be legitimate—if so, it should be explained in the text; if not, see if you can tell what he did. Sometimes the odd figures mean that the experiment has several variables operating at the same time.

Are the diagrams well described and well labeled? Is it clear what they are supposed to be showing? Can you draw other conclusions from them besides the ones drawn by the author? Can you see how he drew his conclusions from them?

3. In work in indexing or classification, is the experimenter operating under illusions regarding bibliographic systems? Does he understand what present systems are like and how they work? Is there any indication that he has ever seen a classification schedule? Does he know about the syndetic part of subject heading? Does he realize the reasons for making authority files or otherwise having exact records? Does his use of terminology show acquaintance with present and past literature? Can you recognize old ideas in new words? Watch particularly for repetition of Ranganathan's ideas.

Does the author realize that the user's question may not be the question he really wants answered? Does he show any familiarity with the basic reference tools in the area covered? Does he think there is only one type of user (the kind who thinks like him)? Does he appear to believe that only one field or area or subject is significant, timely, or paramount? You will think of similar questions as you read.

4. In any report, read the parts describing input especially carefully. Look for such statements as, "Terms were selected . . ." Ask, "By whom? from what source?" This usually means human selection by intuitive means. The human part can be extremely well hidden and you do not find it unless you deliberately look for it.

Look to see what the experimenter did with his input material. Has he altered it, selected certain parts from it by exercising human choice and judgment? What would have happened if he had left it alone? Watch for evidence of human effort in anything called "automatic." Many "automatic" systems are automatically static unless changes or new initial input receive human attention.

5. If the experimenter used a computer, it often means he counted something? What did he count? He may also have compared items and performed a further operation on the basis of the result. Or he may have repeated an operation or series of operations on a definite number of items. What labor did he save by using the computer? What time? When you take away the computer, what does his system do? What does it do that a manual system does not do? What does it not do that a manual system does?

6. If the experimenter used a statistical method, did he use a sample large enough to be valid? Were his selection or sampling methods adequate? If you cannot find out any other way, ask a statistician, actuary, econometrist, psychometrist, or sociometrist. In information science, samples tend to be too limited and too small to be more than a rough indication of what *might* be the case. Think of the size of the field under consideration before you accept any figures. A sampling of 500 Sanskrit scholars could be definitive while a sample of 500 chemists could be quite inadequate.

7. How clearly does the experimenter describe his method? Scientific work is supposed to produce publicly verifiable conclusions. Could you repeat his experiment yourself? Does he want you to take his word or the word of an unidentified "specialist" for a statement of fact? Does he appeal to authority not available to you for verification purposes instead of producing evidence? This is often the case where a value judgment is made.

Does he make comparisons without showing you both sides? For example, the experimenter may say his system works better than "conventional" classification, without making any attempt to show how it is better than, say, Dewey, Universal Decimal, Bliss, Library of Congress, or Colon. This is especially likely if he "proved" all existing classification systems inadequate in a sweeping statement at the beginning of the report. Make the comparison yourself by lining up old and new side by side and demonstrating their treatment of the subject. Surprisingly few experimental classifications can survive this treatment.

Does the experimenter tell you that he has tested some new work of another experimenter and it is not satisfactory, without showing you what he thinks it should have done in order to be satisfactory? Is he using some scale of values which you are not permitted to see?

8. Examine the output. Look carefully at the material from which the experimenter drew his conclusions. Would you draw the same conclusions from this data? Has he given you enough data? Does the experimenter go beyond his data in drawing conclusions? Does his experimental method have application beyond a narrow field? Try the method on another field. Will it produce results in any subject except the one on which he tried it?

9. A large proportion of experiments are poorly conceived, poorly designed, poorly executed and poorly inter-

preted. Can you see anything that the experimenter may have missed? Can you improve on his experiment? Watch particularly for evidence of lack of knowledge and experience in the field the experimenter is covering as this may lead to discovery of operating variables of which he was unaware.

10. Finally, many experimenters in information science were originally "hard" scientists and are not accustomed to writing clear and readable prose. Reduce their sentences to basic English wherever possible. Change the passive voice to the active and put in a subject (usually I). Omit the superfluous verbiage and translate big words to little ones. Most reports turn out to be reasonably simple after such treatment.

Phyllis A. Richmond
Library
University of Rochester

The Training Implications of Automated Personnel Systems

Introduction

Whenever an automated system of any type is installed there will always be training required: training to acquaint employees with the new methods at installation time, and continuing training of the people in the computer department to keep them abreast of the rapid technological changes in automation methods. In computerized personnel systems, in particular, we can identify four major areas of training responsibility. We must train:

1. People whose records are being processed;
2. People who prepare input data;
3. People who run computer systems;
4. People who use computer outputs.

Training People Whose Records are Processed

An automated personnel system from purely a systems designer's point of view is quite like an automatic production status system or a computerized inventory control procedure. But in automated personnel processing each record in the system does not represent a box of hubcaps, it represents a flesh and blood human being. Each punched card can represent a man who may have serious doubts about any machine's ability to properly handle his personnel affairs, or on the other hand, he may have a genuine fear of his statistics being processed by an all powerful "electronic brain."

This fear of the ability of the computer is most dramatically expressed by the enormous increase in voluntary tax return filings received by the Internal Revenue Service since they began to computerize their record keeping.

As a result of over 3,000 personal interviews with a general cross section of the American public, Dr. Robert S. Lee of Columbia University recently wrote that the uneasiness about the computer as an "all knowing super device" causes even more concern than the fear of displacement through automation.

With this as a background, what then is our training responsibility toward the men whose data are being processed by the machines? We must assure each man, that under the automated system two things will happen:

1. He will receive objective and impartial treatment;
2. His treatment will not be impersonal.

We must assure him, for example, that if we plan to search the personnel files for all those men with over 5 years of accounting experience and a reading knowledge of French, that if he has these qualities, he will be selected. Or, that we could impartially locate all men whose absences from the job exceed 15 days per year.

We must assure the man through proper training that even if his processing is done by machine that the contact with him will be by another person who can explain what the computer has produced, how it affects the man, and can make allowances for extenuating circumstances not programmed into the system.

A friend of mine was recently selected by the Army's automated personnel system for overseas assignment. There was no question that this man was due for his overseas rotation at this time, so the computer did its job properly. However, the people in the personnel assignment section delayed his departure for 1 month while his wife recuperated from a serious operation. This is what is meant by objective yet not impersonal action.

The complexity of the system and the number of people controlled will bear directly on the training method, but usually carefully prepared lectures on introductory computer processing, and a clearly written, well-illustrated pamphlet on the system will do the training job. If done properly, they will orient the staff toward a realistic appreciation of what the computer will be doing and will stimulate their cooperation and active support while you are installing the system.

Training People Who Prepare Input Data

The person we are talking about here is the man at the interface between the source of the raw data and the system itself. Specifically, we mean the individual who prepares input forms containing the data that enters the system. This individual may be the same man whose records are processed but usually he is some type of coder, such as a bookkeeper in an accounting office or a clerk in a military pay system.

The reason for the training is that the coder must have a crystal clear understanding of what data to enter into the system, how it is to be represented, and precisely where on the forms it is to be placed. If we can not establish this data entry discipline through proper training then the entire system will fail. Without valid input we have nothing.

There are really only two requirements to insure the effectiveness of the input training effort: (1) Keep the coding instructions simple, and (2) Continually retest and retrain the coders.

The instructions to the coders must be in plain English. We recently assisted a major accounting firm in installing a large cost control system where the instructions to the coder required her to "place the net receipts in domain A, and the gross receipts in domain B" on the coding sheet. All we had to do to reduce the input errors by 4% was to change the word "domain" to the word "position" in the procedures.

In many cases an imaginative systems designer and trainer can convey the coding requirements simply and with very little text by using the decision table technique. A properly designed decision table can easily hold, on one sheet, the information from five or six sheets of written text to quite easily direct the coder to the proper entry.

If many coders are to be trained this is a made-to-order application for the programmed instruction training technique. The method is successfully used by the Internal Revenue Service in training their staff in visual editing of some tax returns, and by a number of computer manufacturers in the training of the coding aspects of computer programming. Programmed instruction texts can be expensive (about 50 hours of writing to each hour of net instruction) but can be a very efficient way of training large groups of coders.

Because the input coding function is so critical a one-time training effort will usually not suffice. Therefore, a routine must be established in the automated system design itself to determine the error frequency by type for each coder so that management can decide when it is economic to recall coders to reinforce their understanding on a data entry problem.

Training People Who Run Computer System

Probably the most obvious training implication of an automated personnel system is the training of the people who will run the system. Here we are concerned with key punchers, tab operators, computer operators, programmers, and systems analysts.

One of the most vital issues here is the question of who will be trained. When any automated system is installed there is usually some displacement of clerical personnel. The extent to which these people can be retrained for responsibilities in the computer department is limited, in my opinion, mainly by the imagination of management, not by the ability of the worker. When the new computer installation is staffed with retrained employees, we have found it to have a far greater chance of success than one staffed by new hires. Aside from the obvious morale advantages, the reason for this is really quite simple. It is usually easier to train someone in computer techniques than it is to teach a man your business.

Computer training is becoming a highly developed art, and if properly trained, a man can be writing productive programs in just a few months. We, for example, over the past 3½ years have trained and placed hundreds of computer programmers who have completed our 25 session, 2 night per week evening course. I dare say that it would take many many times longer than that to train a man in the intracacies of civil service personnel actions.

It is interesting to note that organizations who have for various reasons been forced to follow this retraining policy have found it to be a blessing in disguise. Many banks, with outstanding ADP results, are forced to train from within rather than hire fully experienced programmers who might demand a salary in excess of that of one of the vice presidents. Many progressive railroads hamstrung by union seniority restrictions have found that the manual-operations-experienced man is exceptionally valuable in the computer section.

Training People Who Use Computer Outputs

Probably the most universally neglected training responsibility is that of training the user of the computer outputs, namely, the manager who is expected to act based upon a computer generated report he receives.

It is unfortunate that in our current state of highlv developed third generation software and hardware systems that the training of the executive who prescribes and uses these systems is usually only a stepchild of other training programs. Management training is frequently just a generalization of a computer concepts course for beginning operators or a watered down version of a computer programming course.

We strongly feel that the content of the training for an executive must be something unique to his needs. We recently completed a series of 62 depth interviews into the automation training needs of executives from research organizations, insurance companies, retailers, construction firms, banks, and government agencies. All of the executives were surprisingly consistent in their responses and expressed a need to receive training in two major areas:

1. An ability to recognize the management considerations in computer processing.
2. An ability to react to a management consideration with the appropriate management technique.

More specifically the executives wanted to receive detailed training in the following five areas:

1. How to identify potential systems applications
2. What a manager must know about programming
3. How to prepare a feasibility study
4. How to control the accuracy of data in a computer system
5. How to organize for computer processing.

The list of five areas we mentioned is valid for the state of the computer art as it stands today. It is valid for the needs of personnel record keeping. The future, however, will be sure to expand on this list of training topics when applications such as automated personnel scheduling systems using computer generated simulations are used, or when the growing popularity of the linear programming technique is used in wage and salary evaluation. Both of these techniques have been already successfully applied in these areas.

Summary

In summary then we see that to insure an efficient automated personnel system we must train four major groups of individuals:

1. People whose records are processed
2. Coders who prepare input
3. Those who operate the computer system
4. Managers who use the outputs to make decisions

The effective training of people to work efficiently in a computer environment, in our opinion, represents one of the major challenges to modern personnel administration.

Michael J. Rauseo
Management Research Associates
Arlington, Virginia

Documentation in Thailand

Thailand, which some may know better by its older name of Siam, is still largely an agricultural country, although a considerable amount of industrial development has started in recent years. The value of scientific research as a means of accelerating the country's development is now fully recognized by the government. The first important step taken by the government towards improving scientific research was the establishment of the National Research Council of Thailand as a central body to advise the government on scientific policy.

The National Research Council, in its turn, recommended that the Government should take two further important steps. The first of these two steps was the establishment of a Thai National Documentation Center, to operate a full range of documentation services, including a national science library. The second major recommendation was for establishment of an Applied Scientific Research Corporation of Thailand, which would function as a semigovernmental body for the purpose of setting up and operating national research institutes in the main fields of the applied sciences.

As a matter of fact, one will agree that the National Research Council asked for these two projects in the correct order. It was essential that the initiation of the Documentation Center should procede the expansion of the research activities. Such research work as was already in progress in the country was being severely handicapped by the lack of modern organised facilities for obtaining information. The planning of new research institutes and of their programs, even before they went into operation, would require the services of an efficient documentation center. Only on this condition could the research program be established on a rational basis.

The National Research Council therefore took action to set up a national documentation center. In 1961, on a request from the Government of Thailand, UNESCO started providing specialized aid for this project under the U.N. Technical Assistance program. The aid from UNESCO is mainly in the form of the services of three expert advisers for several years, to assist in planning and implementing the project, and approximately $25,000 in equipment and several fellowships for training the staff of the center.

The Government through the National Research Council, provided funds to the amount of three quarters of a million dollars for the building and equipping of a modern documentation center. This center, the Thai National Documentation Center, went into operation in May 1964, and is at present in a stage of rapid expansion of its services.

The TNDC, to give the center its short title, is designed to carry out a number of functions. It is a national documentation center, providing services of document procurement, bibliography compilation, and translation for science and industry throughout Thailand. It includes a national science library, with approximately 2,000 square meters of stack rooms and reading room. The Center also provides the special library services for the research institutes of the Applied Scientific Research Corporation, which are being established on the same site as the TNDC.

Now to say a few words about the general position regarding scientific library and information services in Thailand, against which the value of this new development, the establishment of the TNDC, can be assessed.

Thailand has at present about 60 scientific libraries, all but one or two of them being located in the capital, Bangkok. They range in size and effectiveness from two or three fairly large libraries with a well-qualified and experienced librarian to small, poorly stocked collections with untrained or part-time librarians. Interlibrary collaboration is better than might be expected, thanks to the existence of an active Thai Library Association, but is still hampered by various factors, including the restrictions that are commonly due to government control. The stock and services of these libraries are usually available to a limited range of users. One or two of the larger libraries provide documentation services, such as procurement of microfilm or photographic copies of papers, to their users; the majority of them, however, do not attempt to go beyond the collection and use of their own stock.

The combined resources of these libraries, even if they were all available for general use, would still be very small. On a reasonable estimate, perhaps one in ten of the published documents that a scientist in Thailand might need to consult would be available in the country. The remaining nine tenths are at present obtainable only by procurement from abroad. In such circumstances, it is obviously not possible to carry out serious scientific work.

In order to improve this situation, the immediate program of the TNDC is to fill the huge gaps by providing a low-cost, fast service of scientific document procurement by modern photographic methods, available to all who can use it. Bangkok libraries are made fully available through this service, since the TNDC has compiled a union card catalogue of periodicals held in these libraries and can supply microfilms or photocopies of any article from them on request. For the remaining 90% of papers, the TNDC has established contact with some 30 documentation centers throughout the world from which microfilm copies can be obtained by the TNDC on request.

This service, together with other supporting services such as bibliography compilation and translation, at least enables the scientist to get on with his work. It is recognized by the TNDC, however, that the full long-term solution requires much more than this.

The TNDC is therefore working in various ways to reduce the dependence of the Thai scientist on time-consuming and expensive procurement of individual papers from abroad. The Center's own library is being stocked as rapidly as possible, particularly with important publications that are not otherwise available in Thailand. Encouragement, help, and advice to other libraries is an avowed part of the policy of the TNDC, since it is realized that the specialized scientist needs the direct services of his own special library in addition to those of the national center. In its training programs for its own staff, the TNDC is recognizing that the keen, carefully selected young people who are now learning the intricate operations of scientific documentation in the Center will one day almost certainly be drawn away to organize and operate specialized centers in the more important branches of science and technology.

The steps described above will, it is hoped, keep the development of scientific documentation in Thailand in its proper position, that is, ahead of the needs of the research workers. The execution of such a program is, of course, not without its problems, but the excellent support of the government is making it possible to solve these as they arise. The more urgent needs, both in the TNDC and in the scientific libraries, are for training facilities, particularly study abroad, and for aid in the massive problem of building up the stocks of libraries to a reasonable level.

We in Thailand are fully aware that collaboration is a two-way operation. We are grateful for the opportunity afforded to us by other countries under which we can draw on their knowledge, their libraries and their services. For our part, we are making every effort to collect and process the scientific literature of Thailand, and we will always be glad to deal with any requests or inquiries in relation to that literature.

A few words must also be mentioned about the recent progress of the Thai National Documentation Center. Practically the TNDC has been established to provide a variety of documentation services both to the staff of the Applied Scientific Research Corporation of Thailand and to research workers and technical personnel in other institutions in Thailand.

The center was officially inaugurated December 2, 1964, and the growth of the demands on the "responsive" services can be seen by comparing the demands on these services during the first and second halves of 1965. The numbers of requests received were:

	Jan.–June 1965	July–Dec. 1965
Document procurement	186	452
Bibliography compilation	20	26
Translation	25	30

The library continues to grow rapidly. The book intake in 1965–2,500 volumes—was equal to the total intake of the previous 3 years of preliminary collection. Research reports, bulletins, etc., received during the year numbered 7,100. The number of scientific periodicals being received by subscriptions, exchange or gift was increased during the year from about 700 to over 1,000 titles, and good progress was made in building up library's holdings of back volumes.

The demands on the microfilming and photocopying services reflected the growing demands on the documentation services. There was also an increasing demand on the the facilities for making copies of scientific documents provided by the user organization. Comparative figures for the two halves of the year 1965, with the annual totals, were:

	Jan.–June	July–Dec.	Total for 1965
Microfilm (frames)	512	18,626	19,138 (3,828 strips)
Photoprints from microfilm	3,897	4,562	8,459
Direct photocopies	800	1,922	2,722
Jobs requested	117	359	476

The services of the printing facilities were also in heavy demand. The offset printing presses carried out jobs totalling 1,065,410 impressions (single-side printing), while the duplicating machines produced 232,260 impressions.

Last but not least, "List of Scientific Reports Relating to Thailand List No. 2" was published in December 1965 with the inclusion of 2,115 items in various field of science and technology.

CHUN PRABHAVI-VADHANA
Thai National Documentation Centre
Bangkhen Bangkok, Thailand

Scientific and Technical Information in Japan*

Mr. Chairman, Ladies and Gentlemen:

It is my pleasure to introduce to you today the general situation in Japanese scientific and technical information. There are many subdivisions within the field of scientific and technical information, so I feel it would be better if I focus on major organizations and their backgrounds and activities in this broad field in Japan.

I have chosen to talk about the Japanese Ministry of Education, the National Diet Library, the Japan Science Council, and the Japan Information Center of Science and Technology.

The Ministry of Education was established as an integral

* Speech given at Science and Technology Information luncheon group on Thursday, May 4, 1967.

part of the Government in 1871. It exercises a profound influence upon the majority of Japanese agencies producing and using scientific information through those of its functions which affect all aspects of education, its control of personnel and policies and appropriations in almost all of the national universities. It is further influential through its subsidizing programs for governmental and private research institutes, academic societies, and through the publications these organizations issue.

The Bureau of Higher Education and Science is one of seven major divisions within the Ministry of Education and this Bureau consists of nine subdivisions. One of these subdivisions is the Scientific Information Section created in August of 1952.

The advisability of establishing a large national information center for science and technology was recognized early—about 1950. In May of 1951 the Japan Science Council, to which I shall refer later, after discussing the matter in a special committee, advised the government to create a large science information center with the further recommendation that this center be developed within the Ministry of Education. The Scientific Information Section was a preliminary to the larger step. However, the necessary budget for a large center was not forthcoming so the Scientific Sections activities continued on a limited scale, with some growth achieved.

The Scientific Information Section launched its work by preparing a union catalog of foreign-language scientific books and periodicals in major Japanese collections. Before that time any one university library, for instance, would have scant information about the collections in other universities.

The Information Section is also charged with broadcasting Japan's scientific achievements abroad, so to speak, and with promoting the international exchange of science information.

The Ministry of Education's most important present contribution to the development of science information activities is its financial and technical assistance to the publication of the *Japan Science Review*. The *Review* is published in three separate sections: (1) Mechanical and Electrical Engineering, (2) Medical Sciences, and (3) Biological Sciences. Each section contains bibliographies and abstracts in its respective field.

The total budget of the Ministry of Education for 1966 amounted to some 700 million dollars, of which 10 million dollars was granted in aid for basic research. Of the latter amount, $200,000 was devoted to the publication of research results, where emphasis was laid on secondary publications with an allotment of $50,000. This was distributed mainly among academic societies to assist them in the publication of periodicals and scientific books.

The second one is National Diet Library. By tradition, libraries in Japan have been regarded as private properties of individuals, organizations, and institutions. The National Diet Library was created in 1948 by Parliamentary Act to serve as a national center of library activities.

Plans for the National Diet Library organization were developed with the advice of leading officers of the United States Library of Congress.

The National Diet Library consists of the Central Library and 33 branch libraries. Branch libraries, which were formerly independent of one another, may be classified into three types. One is the Ueno Branch Library, formerly the Imperial Library. The second type comprises the famous private collections such as Seikado and Toyo. The third type consists of the libraries of government organizations servicing administrative and technical personnel in the government.

Formerly, the administrative headquarters and the major collection of books, journals, and films were housed in the Akasaka Detached Palace, with considerable difficulties, but many of these difficulties have now been overcome by the construction of a new National Diet Library building in late 1961, at the administrative center of the national capital.

The National Diet Library's total volumes numbered above 5,220,000 in 1960.

The Library's budget for book purchasing was about $600,000 in fiscal 1966 of which $500,000 was allocated to the field of science and technology.

The Library offers reference services to Diet members. The Library also serves as a depository for copies of all materials published in Japan. The Library coordinates library activities among government agencies. The Library furnishes general library and bibliographic service to the public, and it promotes international exchange of information, including publishing in foreign languages and participation in international publication projects. The Library also prepares and publishes *Japanese Periodicals Index* of serials published in Japan.

This series contains two editions, a quarterly of social and humanistic sciences, and a monthly of natural sciences. As for the latter, its English edition has been issued since August 1960, with support from the United States National Science Foundation, and 500 copies are being sent regularly to the United States, and 400 to other foreign countries.

Now the Japan Science Council will be explained. Soon after the Second World War, leading scholars in Japan formed a Commission of 108 men under the Chairmanship of Dr. Kankuro Kaneshige, to review overall problems of postwar rehabilitation of intellectual life and organization of research. How could Japan be brought up-to-date in science and technology after the long years of war time isolation?

The Commission recommendations led to the formation of the Japan Science Council as a state organ, with the aims of promoting the development of science and permeating it into the administration and industry, as well as into the life of the nation.

The Council consists of a Secretariat and seven major Divisions: the Humanities; Law and Politics; Economics; Natural Sciences; Engineering; Agriculture; and Medicine and Pharmacology. The members of the Council are elected by their professional colleagues for terms of 4 years. The voting privilege is limited to those who are professionally active and have graduated from a college at least 5 years prior to the elections: Candidates for the Council seats are normally men and women of national distinction in their own intellectual fields.

The Council provides a forum for the exchange of ideas at the highest level, and serves as a liaison agency among many diverse institutions and agencies. It lacks authority to compel acceptance of its advice by the government, but its decisions and recommendations do carry considerable weight. Consequently, the primary function of the Council is that of a policy-making organ. Its secondary function is to stimulate cooperation among all Japanese institutions concerned with science and technology. Its ternary function is to represent the country in international scientific activities.

The scientific information activities of the Council consist of those of the National Committee for Documentation and the Library of the Council.

The National Committee for Documentation, which is a subordinate organ of the Council and the national representative of Japan to the International Federation for Documentation and other international organizations in matters of scientific information.

The Library of the Council, which is attached to the Secretariat, serves the Council. The Library exchanges the publications of the Council with over 70 countries of the world.

To stimulate public interest in the nationwide problem of such activities, the Council carries on a modest but important publication program based on its library resources, which consist of 50,000 books and 60,000 journals, bulletins and so on. It issues discussion papers and it sponsors public lectures and symposia such as one held in April 1959, at which important national representatives discussed control and servicing of scientific information before an audience representing many professions and all regions of Japan.

The last is Japan Information Center of Science and Technology. When the Prime Minister's Science and Technics Agency was established in May 1956, it successfully lobbied in the Diet for support of an organization which it proposed to sponsor; and the JICST came into being in August 1957, as a corporate body contributing to the development of science and technology from foreign and domestic sources.

The organization of the center consists of a President, Vice President, Specialist Committees, a Planning Office, a General Affairs Division and an Information Division, with a Library and a total staff of about 150.

It has a capital fund of about one million dollars and its budget for fiscal 1966 was also about one million dollars.

Operations are conducted in a new four-story building equipped with modern facilities and located advantageously near the principal government offices and the new National Diet Library building.

As a domestic service to Japanese industry, the Information Center provides bibliographies, references, and similar documents in all fields except biological, medical, and agricultural sciences. It translates foreign papers into Japanese upon request, and provides photoreproduction services. It offers the same services to foreign clients, including translation of Japanese papers into English.

The Center also provides abstracting and indexing services. In 1960 about 1,000 Japanese and 2,500 foreign journals, patents, and other publications were abstracted or indexed.

For storage and retrieval of the documents, it generally uses the Universal Decimal Classification System. A specially designed electric computer was installed in May 1961. It uses magnetic tapes for storage and retrieval of scientific information.

The publications of the Center are: *Foreign Patent News; Current Bibliography on Science and Technology;* and the *Center Monthly.*

Foreign Patent News is a weekly, published in Japanese and English regarding chemistry.

Current Bibliography on Science and Technology is issued in a fortnightly series of six parts. They are general and mechanical engineering; electrical engineering; chemistry and chemical industry; geology, mining and metallurgy; civil engineering and architecture; and theoretical and applied physics.

Publications from far countries are collected via air cargo in order to process them in the shortest possible time.

The Center monthly is issued in Japanese to provide information concerning the Center's activities and articles on documentation study.

The Center, in the course of 10 years of operation, has established a nationwide reputation.

The Center has also been an associate member of the International Federation for Documentation since 1957.

All four organizations about which I have told you are governmental or semigovernmental organizations. In other words, even nowadays scientific and technical information business has not yet appeared in Japan.

Finally, I would like to talk about the major inadequate points in this field. They lie principally in the areas of insufficient coordination among relating agencies and in language problems.

One agency still knows little about what others have done or are doing and planning. Of course personnel concerned have been trying to establish close relations with each other, but we feel there are some gaps as well as overlaps in their activities.

I believe most scientists and engineers in Japan can read academic papers in English, but writing papers in English or some other foreign language is a different problem. Scientists and engineers prefer to write in Japanese, to save time. Translation of these papers into various foreign languages adds extra work to the information service agencies.

I am optimistic about the future, however. With the new electric computers we can do anything, eventually. They are a great help even now. And furthermore, the invention of a translating machine, or the new use of computers for that purpose, should lighten the very heavy burden of translating slowly and writing it out by hand.

I thank you very much.

TOHRU KIKUCHI
First Secretary
Embassy of Japan
Washington, D.C.

Letters to the Editor

Dear Sir:

Considerable research has been undertaken in recent years to learn how scientists and technologists keep themselves informed of current work in their fields of interest. The reports of this research emphasize that the most important ways of keeping informed about current work include attendance at international, national, and local professional meetings, the distribution of formal reports and reprints among colleagues, a considerable amount of personal correspondence, and a significant amount of oral communication at meetings. The study of journal articles, the use of abstracting and indexing services, and the use of library and documentation services appear to be of secondary importance in meeting the needs of scientists and technologists. An important reason for assigning these methods secondary importance is the time lag between the current state of research and development activities and the subsequent publication of information about them. This period is variously estimated to be from 1 to 3 years, even though most professional societies continue to exert themselves in attempts to shorten this time lag.

Some ideas for making it possible for scientists and technologists to find information about the current status of research and development are offered below.[1]

In the present age of computerized publication of throw-away annual directories, it should be quite feasible for a professional society to provide more up-to-date information about the activities of its members by adding to the directory a key-word index to the research and development activities of its members. Annual directories of professional societies would then consist of a minimum of three parts: an alphabetical directory of members, a geographical index, and a key-word index to research and development activities.

The key-word index will require a way of gathering information about research and development activities. All professional societies present an annual statement of dues, and while they collect the necessary money from those who

[1] The stimulus for this communication was found in the description of the American Psychological Association's Project on Scientific Information Exchange in Psychology, in the article by Belver C. Griffith and William D. Garvey, "Systems in scientific exchange and the effect of innovation and change." *Proceedings American Documentation Institute,* 1:191-200, 1964.

wish to continue membership as well as members' addresses and official titles, customarily they do not gather much more information. A questionnaire sent with the dues statement could provide the information for a research and development index to the directory. Questions such as these could be asked:

1. Are you currently engaged in research? Describe the title of each research project in no more than 10 words and indicate the starting date and the projected completion date.

2. Are you engaged in a development activity? Describe each activity in not more than 10 words, and include starting date and projected completion date.

3. Have you published reports, articles, books, etc., since last filling out this questionnaire? Please list your formal publications with full bibliographical citations using the following entries as your models. (models omitted here) New members are requested to send a complete list of their publications.

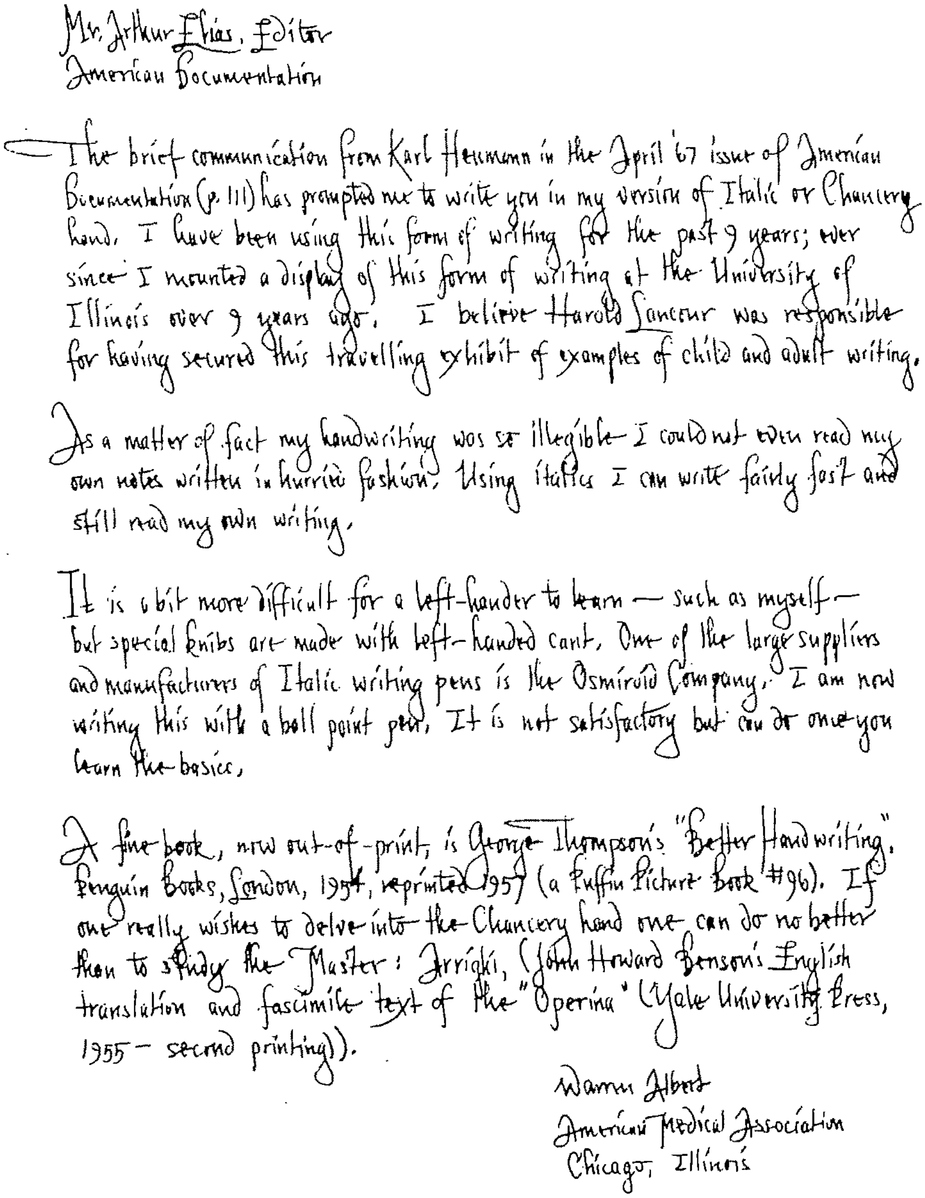
4. Have you prepared manuscripts, informal reports, or any other writings which you would be willing to share with your colleagues? Please list them.

5. Are you willing to send copies of your publications to this society's headquarters so that they may be copied and distributed to all persons who request them? If your answer is "yes," please star the entries above to indicate the items you are sending with this report.

If a professional society does not wish to store and copy publications for sale to interested persons, it could arrange with a library or a commercial service to undertake this activity. For example, the journal articles of *Chemical Abstracts* are kept and sold for the Chemical Abstracts Service by the John Crerar Library in Chicago, and University Microfilms in Ann Arbor stores and photocopies the doctoral theses listed in *Dissertation Abstracts*.

The professional societies which maintain abstracting and indexing services in their particular fields might be willing to match the bibliographical citations obtained from the questionnaires with the citations in their abstracting and indexing services. The entries which are not found in the abstracting and indexing services could be added to the key-word index to research and development activities in the current directories.

C. D. Gull,
Professor of Library Science
Indiana University
Bloomington, Indiana



Mr. Arthur Elias, Editor
American Documentation

The brief communication from Karl Heumann in the April '67 issue of American Documentation (p. 111) has prompted me to write you in my version of Italic or Chancery hand. I have been using this form of writing for the past 9 years; ever since I mounted a display of this form of writing at the University of Illinois over 9 years ago. I believe Harold Lancour was responsible for having secured this travelling exhibit of examples of child and adult writing.

As a matter of fact my handwriting was so illegible I could not even read my own notes written in hurried fashion. Using italics I can write fairly fast and still read my own writing.

It is a bit more difficult for a left-hander to learn — such as myself — but special knibs are made with left-handed cant. One of the larger suppliers and manufacturers of Italic writing pens is the Osmiroid Company. I am now writing this with a ball point pen. It is not satisfactory but can do once you learn the basics.

A fine book, now out-of-print, is George Thompson's "Better Handwriting", Penguin Books, London, 1954, reprinted 1957 (a Puffin Picture Book #96). If one really wishes to delve into the Chancery hand one can do no better than to study the Master: Arrighi, (John Howard Benson's English translation and fascimile text of the "Operina" (Yale University Press, 1955 — second printing)).

Warren Albert
American Medical Association
Chicago, Illinois

Book Reviews

4/67-1R **Theory of Self-Reproducing Automata.** 1966. John von Neumann. Arthur W. Burks, Editor. University of Illinois Press.

John von Neumann made a number of important contributions to the development of modern computers and automata. The present volume has been edited from two of his unfinished manuscripts and so falls naturally into two parts.

In Part I, "Theory and Organization of Complicated Automata," von Neumann presents his views on extremely complicated automata. He begins by discussing computing machines in general and what makes them complex or simple.

Von Neumann went on to discuss rigorous theories of control and information including the statistical information theory of Shannon. He related these results to the theory of computability as formulated by Turing.

It is the last two of the five lectures which are the most interesting. Here, von Neumann discusses the role of complexity in automata. Many similarities between complicated automata and the nervous system are discussed. It is interesting to note that he was intrigued by the numbers of elements involved. There are 10^{10} neurons in the human brain while computers had 2×10^4 tubes when von Neumann gave these lectures. Current computers (IBM 360-91) have roughly 5×10^5 transistors so that natural systems are still much larger than artificial ones. When this material was written, a 1,000 word memory was standard; now 65,000 word memories are common. On the other hand, the memory capacity of a human being is not yet known. Computer components are much faster than neurons.

After considering these questions, von Neumann moved on to discuss the synthesis of complicated automata by other automata. A number of schemes for self-reproduction are mentioned. The possibility of evolution and random mutation is also considered.

In summary, the first part of the book is a semitechnical discussion of the nature of highly complicated automata with consideration of the parallels and differences between abstract and human automata. These lectures are particularly pleasant to read. A serious consideration of these problems requires a knowledge of many areas such as logic, probability theory, information theory, computers, etc. Since von Neumann was competent in all these fields and utilizes them in his lectures, this volume illustrates a first-class mind at work. It is even more impressive when one realizes that we live in a time of pathetic overspecialization.

In Part II, entitled "The Theory of Automata: Construction, Reproduction, Homogeneity," von Neumann set out to construct a self-reproducing automaton. He began by asking five basic questions:

1. When is a class of automata logically universal?
2. Can an automaton be constructed, i.e., assembled and built from appropriately defined "raw materials," by another automaton? Also what class of automata can be constructed by one, suitably given, automaton?
3. Can any one, suitably given, automaton be "construction-universal," i.e., able to construct every other automaton?
4. Can any automaton construct other automata that are exactly like it?
5. Can the construction of automata by automata progress from simpler types to increasingly complicated types? Can this evolution go from less efficient to more efficient automata?

Question (1) has a well-known answer due to Turing. The other questions are all answered affirmatively by von Neumann. This is accomplished by the detailed construction of a self-producing automaton. He starts by postulating a regular two dimensional grid, each cell of which is a 29-state automaton. Using these components, an automaton which is really a (universal) Turing machine is constructed. The constructing machine constructs essentially a tape for a Turing machine and the finite state control unit. The complete machine acts as a (universal) Turing machine. Thus, Question 2 is answered. Question 3 is reduced to Question 2 by giving a plan for converting the constructing automaton into a universal constructor.

Von Neumann reduced Question 4 to Question 3 by showing how to make the universal constructor reproduce itself. The trick is to have a complete description of the constructor on the same tape. Intuitively, this seems impossible since the constructing automaton must contain a complete plan of the constructed automaton and also must be able to understand and execute this plan. This bottleneck can be gotten around by keeping two copies of the information on the tape. One copy is used in the construction: the other copy is passed to the second machine.

Question 5 brings in other issues such as the nature of efficiency. Consequently, this problem is not discussed in the same detail as the others.

In the second part of the manuscript, von Neumann succeeded in answering all of his questions in the affirmative. His detailed construction is easy to follow and quite clever.

The editor, Professor Arthur Burks, has contributed much to the present volume. His many comments are bracketed in the text; this preserves the original flavor and adds explanatory material. The casual reader will find much of this information helpful and most readers will find the chronological material interesting. Advanced readers may find the tutorial comments tiresome and repetitious, but these are easily skipped.

Professor Burks is to be congratulated for doing such a thorough job in finishing these manuscripts. All those people, myself included, who saw the original manuscript can appreciate the work which the editor has done.

Michael A. Harrison
*University of California
Berkeley*

4/67-2R **Symbolic Shorthand System.** 1966. (Rutgers Series on Systems for the Intellectual Organization of Information, Volume VI.) Hans Selye. New Brunswick. 89 pp.

In a brief 10,000 words Dr. Selye and his former librarian, George Ember, here attempt to describe the salient features of the classification system which Dr. Selye has employed in organizing his collection of documents (some 700,000 items, chiefly journal article offprints and photocopies) in the field of endocrinology. This is not a shelf classification, but a classification meant for the organization of a classified catalog. There are some 1,800 "class numbers," distributed abong 20 main classes; the distinctive feature of the system is that the class numbers are made up of mnemonic symbols, which may be combined in various ways according to fixed rules of precedence and order. Thus, the class number for the thyroid gland is

$$\mathrm{Tr}$$

and the class number for lymphomatous thyroditis is

$$\mathrm{Tr\text{-}itis\text{-}Ly}$$

and the class number for antithyroid drugs is

$$\overline{\mathrm{T}}\mathrm{r}\downarrow$$

and if we wished to classify an unlikely article which dealt with the effect of antithyroid drugs on lymphomatous thyroiditis, we would designate the subject as

$$\mathrm{Tr\text{-}itis\text{-}Ly}\leftarrow\overline{\mathrm{T}}\mathrm{r}\downarrow$$

If this same article also dealt with the action of radioiodine in hypothyroidism, this topic could be designated as

$$Tr\downarrow\leftarrow I^*$$

These not unusual examples illustrate several things: the use of truncated syllables suggesting the tissue or the process or the substance designated; the use of symbols such as overlining, underlining, arrows in various directions, asterisks—in all there are 31 signs which are nonalphabetic and nonnumeric. Capitalization or noncapitalization is significant; while "Im" represents infundibulum, "IM" represents immunity or hypersensitivity. Further, as the example shows, citations are posted in as many places in the scheme as the number of topics may demand; they are filed in special "divisions" (*e.g.*, the amino acids tyrosine [Tys] and diiodotyrosine [Dityś] both file in the division of Amino Acids [Amac]); and the element to the right (the agent) determines the order of filing, rather than the element to the left (the target).

This complex but ingenious system has been used by Dr. Selye with great success. Selye is a brilliant man, and has made brilliant contributions to endocrinology and medicine. His published works are supported by lists of references which are astounding in their extent and completeness.

But when all is said and done, the Symbolic Shorthand System remains the *Handapparat* of an individual, whose interests are highly individual. To Selye, endocrinology is not a subject area in the ordinary sense; it is a point of view; it is the place on which he stands and from which he views the rest of the world.

"It should be emphasized," says Selye, "that the system is not necessarily limited to a specific disciplinary field . . . [it is] a system which is applicable to all branches of the life sciences." If by this Selye means that the general notions of the system are adaptable to a field such as rheumatology, *or* such as psychiatry, *or* such as gynecology, to serve the needs of a particular rheumatologist, *or* psychiatrist, *or* gynecologist, then we can agree. But it is inconceivable that the system could be made to embrace at once the entire field of biomedicine, serving the needs of a large number of persons.

Selye wistfully asks "how far could the Symbolic Shorthand System be developed through mechanization"? He says that systems analysts have felt that the system could be mechanized, but he concludes that "the system works so well as it is that the inducement to venture into computerization is not sufficiently attractive." His intuition here is quite sound.

The remaining half of this small pamphlet consists of a "seminar panel discussion," of which 50% is contributed by F. W. Lancaster, formerly associated with the Cranfield Project and now with the National Library of Medicine. Mr. Lancaster, an able man, contributes an interesting discussion of the factors of precision and recall. When he concludes that "theoretically, at least, the system has the capability of a performance range from high recall to high precision" we may be permitted to heave a sigh; but when he says: "In the example, the Order of Predecence says that the effect of adrenaline should precede the effect of cortisone. Suppose, however, we were in an organization in which we were particularly interested in cortisone, or suppose that the emphasis of the document expressed cortisone. Then there is no reason why we could not reverse the order and produce our own weighting system to suit our own needs"—then we may be permitted to oppose thumb to forefinger.

Those seriously interested in exploring the structure of this system should consult *Symbolic Shorthand System (SSS) for Physiology and Medicine,* by Hans Selye and George Ember, 4th ed., Montreal, 1964 (xxxvi, 238 p.), $7.90.

FRANK B. ROGERS
University of Colorado Medical Center

4/67-3R **A Checklist for the Organization, Operation and Evaluation of a Company Library.** 2d rev. ed. 1966. Eva Lou Fisher. Special Libraries Association, New York. 61 pp.

The nonlibrarian management or library administrator will find, in Miss Fisher's checklist, nearly all of the questions that should be asked to analyze the library requirements and services to be provided in a company environment. Following each major question are a list of references published, with few exceptions, between 1960 and 1966 which have pertinence to the question.

Part I raises 12 general problems of management. The reader may follow the cross-references included in the text should he desire more detail or allied information on any topic.

Part II covers 26 specific problems of library operations from Acquisition to Statistics and for the person who wants to start a small, new library there is Part III "Where to Start."

The concept and general approach is excellent. The booklet is an invitation to consider the major problems of management and operations and, if the invitation is accepted, directs the reader to the literature. The checklist is, indeed, a valid contribution to the literature of librarianship. This fact does not mean the compilation is the complete and total answer on how to organize operate, and evaluate the company library.

It may be that the basic organization of the text should have been revised and updated as were the references. An increasing attention is being given to the use of computers in libraries. This is reflected in a number of the papers cited but is handled only in passing in the text.

The correlation between the text and the references cited could be strengthened. In the section on the form of library organization, the articles on "Planning the New Library" are cited. In most cases these refer to the physical layout rather than the organizational structure. This reviewer would prefer having them appear in Section X of Part II where the topic is "Space." The practice of cross referencing does not lead, in this case, the reader of Section X to the article cited under General Problem 3.

The extensive list of citations in some sections and the paucity of them in other sections suggests there are areas in library literature which are not covered as well as they should be. Library journal editors might review the references in the checklist and other general texts and develop guide lines for suggesting possible subject areas to their writers.

The failure to use boldface type for the margin headings in Part I and the lack of consistency in underlining make the use of this section a little more difficult than for Part II and III.

G. E. RANDALL
Research Library, IBM
Thomas J. Watson Research Center

4/67-4R **Anglo-American Cataloging Rules.** 1967. Prepared by the American Library Association, The Library of Congress, The Library Association and the Canadian Library Association. North American text. American Library Association, Chicago.

A dictionary defines *Rules* as principles regulating the procedures or methods necessary to be observed in the pursuit or study of some art or science. Cataloging rules, therefore, could be defined as principles regulating cataloging procedures or methods. *Principles,* on the other hand, are defined as fundamental assumptions forming the basis of a chain of reasoning. To understand cataloging rules we must, therefore, understand the principles upon which the rules are built. Catalogers, not unlike other librarians, seem to be extraordinarily uninterested in principles upon which their professional work is based. It is safe to presume that most of our activities are based upon countless questionnaires, ad hoc assumptions, and almost always on the present-day conditions or individual experiences in the individual library. The best illustration of my point can be found on page vi of the new *Anglo-American Cataloging Rules:*

> It is regrettable that, because of the great size of many American card catalogs, it was necessary . . . to agree . . . that certain incompatible American practices be continued in the present rules.

The above statement is particularly difficult to understand when one considers the illustrious roster of names that decorates the present rules: Seymour Lubetzky, the originator and prime promoter of the new rules and also the most outspoken advocate of "principles" was the *Anglo-American Cataloging Rules'* first editor (1956-62); C. Sumner Spalding was the Rules' second editor (1962-66); Wyllis E. Wright was the Chairman of the Catalog Code Revision Committee; P. S. Dunkin was a member of the Steering Committee; Ruth C. Eisenhart was a consultant (1961-64); and Richard Angell was a member of the General Committee.

All six of them were members of the seven-man official American Delegation to the International Federation of Library Associations' International Conference on Cataloging Principles in Paris in 1961. During the Conference, the American delegation agreed to all but two votings (conference's principles 10.3 and 12). In spite of this, the present rules departs from the International Principles (Principles: 9.12, 11.14, 9.4, 9.5; new code pp. 3-4).

This departure would have been justified if the demands of computerized processes were making it obligatory but "the problems of machine arrangement of entries in automated systems were not ignored but no action could be taken" (p. vi). The new *Rules* therefore are neither purely "international" or "modern," nor are they strictly based upon principles. They are at their best well-edited, topographically improved, and pleasantly compiled traditional cataloging rules.

Upon close examination, we find that the stated principles are so often contradicted that instead of rules based upon principles we have again a cataloging code based on practices. The entry of the work, for example, is based upon the statements that appear:

1. On the title page
2. On any part of the work
3. In the first work in the collection
4. In the first edition of the work
5. In related work

When there is suspicion or evidence that the statements are erroneous or fictitious the entry is based on reference sources or on the consensus of scholarly opinion.

In case of joint authorship the work can be entered under:

1. The first mentioned on the title page
2. The author given topographic or wording prominence

or

3. the author whose heading is first in alphabetical order

In further analysis we discover that not only the responsibility of the author or representation on the title page have to be taken into account but also:

1. Formal history (as in case of corporate authorship)
2. Official approval of the institution
3. Country of the library for which the cataloging is done
4. English language
5. Library of Congress practices

Not only the principles contradict each other but also the examples add to the confusion. The rule "enter under title a work that is of unknown or uncertain authorship" is illustrated by:

> La capucinière; ou, Le bijou enlevé a la course. Poème (possibly by Pierre François Tissot; erroneously attributed to Pierre Jean Baptiste Nougaret)

Two examples later, the following occurs:

> A true character of Mr. Pope
> (author uncertain, generally attributed to Jean Dennis)
> Main entry under Dennis.

Also entered under title are:

> Bibliography of agriculture. National Agricultural Library.
> and Who's Who in designing . . . Members of. International Association of Clothing Designers.

> *but* Biographical directory of the American Political Science Association . . . Edited by Franklin L. Burdette. Washington, D.C. American Political Science Association.
> and Buyers' guide, British Jeweler's Association.
> are entered under the associations.

Form of name also breaks the stated rules. The following are selected examples:

> Philip II, King of Spain *not* Felipe II, King of Spain

> *but* Jan III Sobieski, King of Poland *not* John III Sobieski, King of Poland

> Catherine II, Fmpress of Russia, *not* Ekaterina II, Empress of Russia and Nasser, Gamal Abdel *not* Abd al-Nasir, Jamal

> *but* Evtushenko, Evgenii Aleksandrovich *not* Yevtushenko, Yevgey and Staravinskii, Igor Fedorovich *not* Stravinsky, Igor

> Disraeli, Benjamin, Earl of Beaconsfield *not* Beaconsfield, Benjamin Disraeli, Earl of

> *but* Newcastle, Margaret Cavendish, Duchess of *not* Cavendish, Margaret, Duchess of Newcastle

> Palestrina, Giovanni Pierluigi da

> *but* Giovanni da Ravenna

Corporate bodies have their equal amount of inconsistencies. Thus:

> Unesco *not* United Nations Educational, Scientific and Cultural Organization and Euratom *not* European Atomic Energy Community

> *but* North Atlantic Treaty Organization *not* NATO

> Loyola University, Chicago *and* Newman Club, Brooklyn College

Names of libraries follow the same pattern:

> Bibliothèque nationale (France)
> Rio de Janeiro, Biblioteca Nacional
> National Agricultural Library
> Kongelige Bibliotek
> British Museum
> U.S. Library of Congress
> etc.

There are also some beautifully reassuring statements: "17 B. Works not of corporate authorship. If the work would not be entered under corporate body under the provisions of A above or if there is doubt as to whether it would, enter it under the heading under which it would be entered if no corporate body were involved. Make an added entry under the body unless if functions solely as publisher."

We should also mention new spelling rules: Muhammad, the prophet (rule 27 A) instead of Muhammed, the Prophet.

In spite of the above, there are basic improvements of the *Anglo-American Cataloging Rules* over the *ALA*

Cataloging Rules for Author and Title Entries (1949). First and above everything else, the new *Rules'* inclusion of the rules for descriptive cataloging is a welcome innovation. The descriptive rules follow the long accepted Library of Congress rules. They are, however, greatly expanded and brought up to date. Chinese, Japanese, Korean, Hebrew, Russian, and Yiddish examples are welcomed.

The 153 pages devoted to the problem of descriptive cataloging are probably the best part of the total rules.

The glossary, capitalization, abbreviations, numerials, punctuation, and diacritics are equally useful. The Index, however, is very uneven: definition of author has two references to almost identical statements while Festschriften have only one reference (to contents) omitting references to entry (33H) and the like.

In conclusion, we can say that our new rules, although ignoring the most pressing needs for international standardization, full implementation of generally accepted principles, and neglecting computerized processing are just as fine as the former rules were at the time of their publication. The cataloger does not have to measure any more the distance to the cemetery—he still must however go outside the city limits and see if the church is "in the open country" (rule 98c). The cataloger now knows how to deal with spirits, spiritual media, and spiritual communications (rule 13C).

Andre Nitecki
Assistant Professor
School of Library Science
Syracuse University

4/67-5R **Information Retrieval with Special Reference to the Biomedical Sciences.** 1966. Wesley Simonton and Charlene Mason, Editors. University of Minnesota, Minneapolis. 199 pp.

This softbound brochure contains the 14 papers and corresponding discussions from the Second University of Minnesota Library School Institute on Information Retrieval, Minneapolis in November 1965 (the first, "Information Retrieval Today" was held in September 1962). In an introductory paper on "Patterns and Problems" Dr. Maurice Visscher of the Physiology Department, University of Minnesota Medical School cites at length from oincial science information studies and approvingly quotes Richard Orr. The state of art of mechanized indexing is cogently discussed by Mary Elizabeth Stevens of the National Bureau of Standards stressing the "rediscovery" aspect of our present efforts: the sequential-step camera discussed at a library conference in New York City in 1853, or the key word-in-context (Chem. Titles, etc.), index developed independently by Luhn and Ohlman around 1958 and later traced by the latter at least a hundred years back to British librarian Andrea Crestadoro's word-in-title index for the British Museum in the 1850's. Miss Stevens also felicitously discourses on differences between "machine-" and "people-indexing," in the context of researchers by Don Swanson, John O'Connor, Tukey and others. M. M. Kessler of MIT discusses search strategies of his project MAC in some detail; Norman Shumway of the National Library of Medicine reviews vocabulary construction MEDLARS subject headings; Louise Darling of the UCLA reports on their MEDLARS experiences; Honeywell 800 to 7094, etc., describing the procedure in detail. Dr. Joseph Izzo of the University of Rochester discusses *Index-Medicus* coverage and identification of diabetes-related literature; a counterpoint on the same theme by Dr. Arnold Lazarow of the University of Minnesota follows; what price a trialogue between Honeywell 800, GE 225 and a Control Data [illegible]? The "profile" of diabetes literature, drawn in [illegible] charts and 9 tables, by Elmo Brekhus, an associate of Dr. Izzo at the University of Minnesota, appears under the unrevealing title of "Newer Methods of Document Handling." Jacqueline Felter of Union Catalog of Medical Periodicals, N. Y. Medical Library Center writes entertainingly and candidly about programming problems and solutions in preparing a computerized Union list of holdings of medical periodicals in Greater New York libraries; while Evelyn Moore, University of Washington, St. Louis, discusses their computerized serials control and book circulation. Frederick Kilgour of Yale University Library outlines in rather general terms the "Basic Systems Assumptions of the Columbia-Harvard-Yale Medical Libraries Computerization Project," while Mrs. Henriette Avram of Library of Congress discusses the card-catalog computerization at LC; Dr. M. M. Cummings of the National Library of Medicine presents a general plan for development of the medical libraries, and Foster Mohrhardt of the National Agricultural Library concludes with a discussion of the "National Information Systems."

It is not entirely clear for whom this volume is intended, in addition to being presumably distributed to all registrants at no extra cost. The recent proliferation of published proceedings of conferences and colloquia, sessions and symposia, on various aspects of biomedical communication and documentation, has already arrogantly taken so much valuable shelf space, that it is difficult to enthuse over another ¾" thick side-stitched tome of 200 single-spaced elite-typewritten 9½"×11" pages, especially as the absence of any index leaves the reader with no other easy pathway through the tome than a one-page table-of-contents listing of titles and authors. A list of registrants in this meeting, and their identification in the question-and-answer sections following each paper would have improved this book, because these sections form an articulate and lively contrast to occasional pomposities in the papers themselves. Virtually all of the charts, tables and specimen printouts accompanying the presentation by Shumway, Izzo, Lazarow, Brekhus, Felter, Moore, and Avram, are well-chosen and useful reference material. Conceivably, any critical comments about the book are attempts to crash an open door; perhaps editor Dr. Simonton modestly felt that the publication was not deserving of general distribution (no price is quoted). Yet there is much of permanent reference value here for biomedical documentalists; certainly not less than in some of the much more expensively printed hardbound books put out by well-known publishers on the same general subject over the past 10 years. I have little doubt but that the volume will find and hold its place among its many competitors on the overflowing biomedical documentation reference shelf.

Boris R. Anzlowar
Pharmaco-Medical Documentation
Chatham, New Jersey 07928

4/67-6R **Coordinate Indexing.** 1966. John C. Costello, Jr., Graduate School of Library Service, Rutgers, The State University (Rutgers Series on Systems for the Intellectual Organization of Information Vol. VII), Edited by Dr. Susan Artandi. Supported by the National Science Foundation. New Brunswick, New Jersey. 218 pp.

The objectives of this series are stated in the Preface as follows: "The investigation is intended to examine the various methods or systems individually, study them in depth within the framework of a seminar series, and then produce a group of papers which, in addition to being state-of-the-art contributions to the scholarship of the field, should also serve as a basis for the ultimate objective, systems comparisons. Each paper then should be a description, a discussion, a critique, a collection of facts and data."

Evaluated against the above criteria, this particular volume falls well short of its goals. It fares best as a *description* of coordinate indexing systems and the methods and procedures associated with them. In this capacity it is excellent and probably no better single how-to-do-it book exists on the subject. As a discussion in the seminar framework, however, it is terribly overbalanced toward the source paper (185 pages or 96% of the actual text) and away from the comments of the attendant panel of experts—Giuliano, Warheit, and Bernier (9 pages or 4% of the text). The contributions of the panel are well worth reading but are so fragmentary that they merely whet the appetite.

The author does point out most of the well-known strengths and weaknesses of coordinate indexing systems. The way in which he does this, however, hardly qualifies as a critique, at least in the scholarly sense of the word. Appended to the volume is an extensive bibliography of 139 references, only one of which is, I believe, referred to specifically in the text. There is an almost complete absence of utilization of the literature or of study results from other workers. There is no hard data on how devices such as roles, links, and weights have fared in tests. There is no data on how much more time it takes to index using such devices. Generalizations which should cite some experimental or factual support are commonly offered to the reader presumably to be taken on faith (Example: p. 186, par. 3). Time and again remarks demanding a reference, a source, a footnote, or some of the other paraphernalia necessarily and justifiably associated with state-of-the-art studies are ignored, leaving the reader, or this reader at least, rather uneasy.

The intent may have been to respond somewhat to the call for a "collection of facts and data" through the insertion of Appendix 2, listed in the Table of Contents as a "Summary of Data for Five Operating Coordinate Indexing Systems." This Appendix was, however, missing from the review copy and no volumes containing Appendix 2 could yet be located at the time of this writing.

What the book does it does logically, thoroughly, and with an ordered approach that organizes a great deal of material for the reader. Alternative working methods and approaches are described in exhaustive detail. The delineable steps in standard procedures are outlined fully and expertly. The author does an excellent job of providing definitions, collecting synonyms, and indicating other terminological problems or confusions that exist in the field. However, after listing a group of synonyms he quite frequently fails to select one for his own use and continues to run through the entire string every time he uses the concept (Example: p. 71, par. 3).

This repetitive, tutorial approach, fine for the classroom but inappropriate here, plagues the book throughout and makes it at least a good 25 pages longer than it need be. Some examples of extensive repetitive passages are in order:

1. Definition of "Related Terms" (p. 94, 170-171).
2. Coordinate indexing can be analytical, or clerical, or various combinations of both (p. 20, 28, 31, 45).
3. Data concerning the document's physical or bibliographic characteristics are of use as well as data concerning its content (p. 17, 57).
4. Generalists make better indexers than specialists (p. 47, 82).
5. Indexers must be familiar with the subject matter (p. 32, 82).

It is difficult to tell whether the redundancy of the text is to be attributed to the author's background in developing instructional manuals and syllabi for the Battelle course on Coordinate Indexing or to the fact that he was apparently required to follow a Rutgers-provided outline not of his own construction.

The author's personal preferences come across keenly; for example, his very strong bias towards subject-qualified indexers is apparent throughout the paper. This frequently relates to a negative bias against machine indexing, as in the following statement: "The accomplishment of coordinate indexing by clerical personnel, or by other personnel not professionally qualified to comprehend what the document actually discusses, can be considered as ineffectual as machine indexing" (p. 30). Machine indexing, as described by the author, is, however, limited to frequency count indexing. The reader will look in vain for any reference in the text to automatic indexing studies or to the fact that they may involve the use of criteria other than straight frequency counts.

Though generally good on terminological problems, the text occasionally makes the mistake of using a specialized term quite a long time before it gets around to defining it. Some examples are "terminal-digit card," "rdl card," "Recall," and "Relevance." This could cause many readers some difficulty.

The four main sections of the text, and their respective page lengths, are as follows: Input (90 p), Store (37 p.), Searching (17 p.), and Output (16 p.). As can be seen, these sections get progressively smaller; relatively speaking, there is also a falling off in thoroughness and quality as one proceeds.

In the section on the Store one finds the following statement: "To the extent that microimage chips and film reels are suitable devices for storage of coordinate indexes, they should be used only for static or historical documents which will require no change to the stored images" (p. 124). It may be of interest that this statement is probably already obsolete in view of the development of certain new equipment, the use of which has just been reported in *American Documentation*.[1]

In that part of the Searching section dealing with the pros and cons of "generic posting" (p. 166-169), the author fails to observe that a computer retrieval system can be designed so that the searcher has the *option* of searching *automatically* on the terms "narrower" or "broader" to the search terms he has initially selected. Such a system would obviate any need for "generic posting" and would save a great deal of storage space.

The section on Output leaves perhaps the most to be desired. Among other things the author says, "Links and roles, as they have been defined here, have been used in unit concept indexing only since 1961. As yet, there has been insufficient experience with their use in retrieval to permit the preparation of more than a handful of reports on their effectiveness; and those which have been published reflect the availability of something less than adequate substantiating data" (p. 190). Considering the claims that have been made for these devices over nearly the entire previous text of the book, most readers will, I am sure, feel somewhat dismayed by this late announcement. The "handful of reports" are not identified.

The author closes with a series of "Indeterminacy Principles," for which he makes the claim that "the very acknowledgment of their existence negates the value of statistical exercises in the name of research on relevance ratio and recall ratio" (p. 193). Reciting over and over again the imperfections of perception and communication, like a kind of litany, he writes himself into such a neat solipsistic corner that one is surprised to find him contributing to the panel discussion which follows.

As a physical object the book is sturdy and well bound. Typographically it is mediocre, with section and subsection headings not underlined, all-capitalized, or boldfaced, and therefore melting hopelessly into the rest of the typed text. As mentioned previously, Appendix 2 seems to have been accidently left out, at least in the copies examined. There is no index. Typographical errors appear slightly more frequently than one is willing to overlook. A list of some 20 such quibbles, noticed casually, without special searching, is being sent to the editor. Oddly, Dr. Giuliano's name appears incorrectly as "Guiliano" consistently throughout the report of the Panel Discussion, but correctly elsewhere.

W. T. Brandhorst
Documentation Incorporated
Bethesda, Maryland

[1] Kozumplik, W. A. and Lange, R. T. Computer-Produced Microfilm Library Catalog. *American Documentation*, 18: 67-80 (1967).

ADI Chapters and Secretaries

CENTRAL OHIO CHAPTER

Miss Arveta McKim
Chemical Abstracts Service
Ohio State University
Columbus, Ohio
614-293-5022

CHICAGO CHAPTER

Miss Patricia Llewellen
IIT Research Institute
10 West 35th Street
Chicago, Illinois 60616
312-225-9630

DELAWARE VALLEY CHAPTER

Miss Marilyn Leasure
Technical Library—Louviers
E. I. du Pont de Nemours & Co.
Wilmington, Delaware 19898
302-366-4242

INDIANA CHAPTER

Mr. Asa N. Stevens
6157 E. St. Joseph Street
Indianapolis, Indiana 46219
317-357-6460

LOS ANGELES CHAPTER

Sister Mary Lucille
Dean, School of Library Science
Immaculate Heart College
2021 N. Western Avenue
Los Angeles, California 90027
213-462-1301, ext. 297

METROPOLITAN NEW YORK CHAPTER

Miss Betty Jean Dougherty
Port of New York Authority
111 8th Avenue
New York, New York 10011
212-620-7000

NEW ENGLAND CHAPTER

Miss Virginia Valeri
Arthur D. Little, Inc.
15 Acorn Park
Cambridge, Massachusetts 02140
617-864-5770

NORTHERN OHIO CHAPTER

Miss Helen Skowronska
Sherwin-Williams Co.
P.O. Box 6027
Cleveland, Ohio 44101
216-TO1-7000

PITTSBURGH CHAPTER

Mr. James Brandt
ALCOA Research Laboratories
Box 772
New Kensington, Pennsylvania 15068
412-337-6541

POTOMAC VALLEY CHAPTER

Mrs. Joan Mavity
Herner & Co.
2431 K Street, N.W.
Washington, D.C. 20037
202-965-3100

SAN FRANCISCO CHAPTER

Mrs. Anne Raphael
176 Osage Avenue
Los Altos, California 94022
415-323-6138

SOUTHERN OHIO CHAPTER

Mrs. Esther Norton
6061 Crittenden Drive
Cincinnati, Ohio 45244
513-684-3111

SOUTH TEXAS CHAPTER

Dr. Kenneth Griffith
M.D. Anderson Hospital and Tumor Institute
6723 Bertner Drive
Houston, Texas 77025
713-JA9-4311

UPSTATE NEW YORK CHAPTER

Mrs. Pauline Atherton
School of Library Science
Syracuse University
Syracuse, New York 13210
315-476-5571, ext. 3823